Monocular 3D Human-Environment Understanding: From Interaction to Reconstruction

Zhi Li

A dissertation submitted towards the degree
Doctor of Engineering Science (Dr.-Ing.)
of the Faculty of Mathematics and Computer Science
of Saarland University

Saarbrücken, 2026

Date of Colloquium:	08.05.2026
Dean of the Faculty:	Prof. Dr. Jan Reineke
Chair of the Committee:	Prof. Dr. André Brinkmann
Advisor:	Prof. Dr. Bernt Schiele
Reviewers:	Prof. Dr. Bernt Schiele
	Prof. Dr. Gerard Pons-Moll
Academic Assistant:	Dr. Wolfgang Stammer

Abstract

Understanding and reconstructing 3D human and environment from monocular observations is a fundamental yet profoundly challenging problem in computer vision. Without stereo or multi-view information, monocular systems must infer depth, motion, and spatial structure from inherently ambiguous visual cues. However, the ubiquity of monocular cameras in autonomous systems, robotics, and consumer devices makes this setting not only practical but also essential. This thesis explores a unified framework for monocular 3D human-environment understanding that progresses through a series of self-supervised or weakly-supervised models, moving from capturing human motion to adapting to changing environments and ultimately reconstructing them.

The first part investigates how environmental cues can be exploited to improve human motion understanding from monocular inputs. Specifically, physical constraints—such as ground contact, support, and body-environment proximity—are leveraged to guide pose estimation. A factorised correction-based framework is proposed for multi-person monocular 3D pose estimation, enabling stable optimisation over imperfect initial predictions. Based on this foundation, a contact-guided motion capture method is introduced, sampling from pose manifolds while enforcing dense contact consistency with the scene. These methods demonstrate how even limited monocular information can be enriched through structured interaction with the surrounding environment.

Beyond human-centric modelling, the next stage examines how human motion itself can be used to recover environmental changes. In dynamic or deformable settings, static scene assumptions no longer hold. To address this, a joint reconstruction framework is developed to simultaneously estimate 3D human motion and environment deformations from monocular video. This approach captures mutual influence: humans adapt to the scene, and their movements reveal the scene's pliability and evolution. Grounded in optimisation, this formulation models environment deformation through human motion, providing a pathway toward high-fidelity dynamic scene reconstruction.

As scenes evolve—both spatially and across domains—monocular systems must remain robust to distribution shifts. To address this, a source-free test-time domain adaptation framework is proposed for monocular depth estimation. A self-supervised optimisation strategy is employed to adapt depth predictions to unseen target domains during inference, without access to source domain data or annotations. By leveraging geometric consistency and photometric cues available at test time, this method effectively mitigates domain shifts commonly encountered in outdoor driving scenarios. Unlike prior approaches that require offline retraining or access to labelled source data, this solution is plug-and-play, efficient, and enhances generalisation in a fully unsupervised setting.

The final stage turns toward full scene reconstruction from a single view. Methods are developed for semantic 3D occupancy prediction from monocular images, enabling feed-forward single-frame inference without reliance on ground-truth occupancy or LiDAR supervision. The approach begins with a NeRF-based volumetric rendering formulation to align 3D semantic predictions with 2D annotations through differentiable rendering losses. Within this framework, a multi-task interaction strategy is specifically designed to improve the synergy between semantic supervision and geometric reconstruction. By integrating semantic and geometric reasoning in a unified formulation, this method enables rich 3D scene understanding with minimal supervision. Despite being trained only with 2D supervision, the system can recover meaningful volumetric

structure from single images, offering a practical step toward self-supervised monocular 3D reconstruction.

Across these stages, the contributions in this thesis form a coherent progression toward robust, self-supervised 3D perception from monocular visual input. From capturing interaction to reconstructing structure, the presented framework demonstrates how machines can perceive and interpret the 3D world through the narrow lens of a single camera—without requiring expensive sensors or annotations. This opens new possibilities in dynamic scene understanding, human-centric computing, and embodied AI.

Zusammenfassung

Das Verstehen und Rekonstruieren dreidimensionaler Mensch-Umwelt-Beziehungen aus monokularen Beobachtungen stellt eine grundlegende, jedoch äußerst anspruchsvolle Herausforderung in der Computer Vision dar. Ohne Stereo- oder Multiview-Informationen müssen monokulare Systeme Tiefe, Bewegung und räumliche Struktur aus von Natur aus mehrdeutigen visuellen Hinweisen ableiten. Die allgegenwärtige Verbreitung monokularer Kameras in autonomen Systemen, der Robotik und in Konsumgeräten macht dieses Szenario jedoch nicht nur praktikabel, sondern auch essenziell. Diese Dissertation untersucht ein einheitliches Rahmenwerk zur monokularen 3D-Erfassung von Mensch-Umwelt-Beziehungen, das sich schrittweise durch selbstüberwachte oder schwach überwachende Modelle entwickelt – von der Erfassung menschlicher Bewegung bis hin zur Anpassung an sich verändernde Umgebungen und deren Rekonstruktion.

Im ersten Teil wird untersucht, wie Umweltinformationen genutzt werden können, um das Verständnis menschlicher Bewegungen aus monokularen Eingaben zu verbessern. Insbesondere werden physikalische Einschränkungen – wie Bodenkontakt, Stützflächen und die Nähe zwischen Körper und Umgebung – zur Steuerung der Posenabschätzung herangezogen. Ein auf faktorisierter Korrektur basierendes Framework für die monokulare 3D-Pose-Schätzung mehrerer Personen wird vorgestellt, das eine stabile Optimierung auf Grundlage unvollständiger Anfangsschätzungen ermöglicht. Aufbauend darauf wird eine kontaktgeführte Motion-Capture-Methode eingeführt, die aus Posenmannigfaltigkeiten sampelt und dabei dichte Kontaktkonsistenz mit der Szene durchsetzt. Diese Methoden zeigen, wie auch begrenzte monokulare Informationen durch strukturierte Interaktion mit der Umgebung erweitert werden können.

Über die rein menschzentrierte Modellierung hinaus wird im nächsten Abschnitt untersucht, wie menschliche Bewegungen zur Erfassung von Umweltveränderungen genutzt werden können. In dynamischen oder verformbaren Szenen gelten statische Annahmen nicht mehr. Hierzu wird ein gemeinsames Rekonstruktions-Framework entwickelt, das die 3D-Bewegung des Menschen und Umgebungsverformungen aus monokularem Video gleichzeitig schätzt. Der Ansatz erfasst die wechselseitige Beeinflussung: Menschen passen sich an die Umgebung an, und ihre Bewegungen offenbaren deren Formbarkeit und Entwicklung. Auf Optimierung basierend, ermöglicht dieses Verfahren die Modellierung von Umgebungsverformungen durch menschliche Bewegung und eröffnet einen Weg zu hochpräziser Rekonstruktion dynamischer Szenen.

Da sich Szenen sowohl räumlich als auch domänenübergreifend weiterentwickeln, müssen monokulare Systeme robust gegenüber Verteilungsverschiebungen bleiben. In diesem Kontext wird ein source-free Testzeit-Domain-Adaptations-Framework für die monokulare Tiefenschätzung vorgeschlagen. Eine selbstüberwachte Optimierungsstrategie wird eingesetzt, um Tiefenvorhersagen während der Inferenz an unbekannte Ziel-Domänen anzupassen – ohne Zugang zu Quell-Daten oder Beschriftungen. Durch die Nutzung geometrischer Konsistenz und photometrischer Hinweise zur Testzeit kann dieser Ansatz effektiv Domänenverschiebungen kompensieren, wie sie häufig in realen Fahrszenarien auftreten. Im Gegensatz zu früheren Methoden, die Offline-Neutraining oder beschriftete Quelldaten erfordern, ist dieser Ansatz plug-and-play-fähig, effizient und verbessert die Generalisierungsfähigkeit in einem vollständig unbeaufsichtigten Setting.

Die letzte Phase dieser Arbeit widmet sich der vollständigen Rekonstruktion von Szenen aus einer Einzelansicht. Es werden Methoden zur semantischen 3D-Occupancy-Vorhersage auf Basis monokularer Bilder entwickelt, die eine vorwärtsgerichtete Einzelbildinferenz ohne Bodenwahrheiten oder LiDAR-Supervision ermöglichen. Der Ansatz basiert auf einer NeRF-basierten

volumetrischen Rendering-Formulierung, bei der 3D-sematische Vorhersagen mittels differentieller Renderingverluste an 2D-Annotationen angepasst werden. Innerhalb dieses Frameworks wird eine Multi-Task-Interaktionsstrategie gezielt entworfen, um die Synergie zwischen semantischer Supervision und geometrischer Rekonstruktion zu verbessern. Durch die Integration semantischer und geometrischer Schlussfolgerungen in einem einheitlichen Modell ermöglicht die Methode ein umfassendes 3D-Szenenverständnis bei minimalem Supervisionsaufwand. Trotz der ausschließlichen Verwendung von 2D-Supervision gelingt es dem System, aus Einzelbildern bedeutungsvolle volumetrische Strukturen zu rekonstruieren – ein praxisnaher Schritt in Richtung selbstüberwachter monokularer 3D-Rekonstruktion.

Insgesamt ergeben die Beiträge dieser Arbeit eine kohärente Entwicklung hin zu robuster, selbstüberwachter 3D-Wahrnehmung aus monokularer visueller Information. Vom Erfassen von Interaktion bis zur Rekonstruktion von Struktur wird ein Framework präsentiert, das Maschinen ermöglicht, die dreidimensionale Welt durch die enge Linse einer einzelnen Kamera zu interpretieren – ohne den Bedarf teurer Sensorik oder aufwendiger Annotation. Dies eröffnet neue Perspektiven für dynamisches Szenenverständnis, menschzentriertes Rechnen und verkörperte künstliche Intelligenz.

Acknowledgements

Undertaking a doctoral degree is a long journey, marked by periods of discovery, challenge and quiet perseverance. It is a path that cannot be travelled alone, and the completion of this dissertation owes much to the generosity, insight and steady presence of many individuals. I would like to take this opportunity to express my gratitude to all those who, in different ways, have shaped and supported the work presented here.

I would first like to extend my deepest thanks to my supervisor, Professor Bernt Schiele, whose guidance, clarity and principled support have been a defining influence throughout my PhD. His unwavering commitment to academic excellence and his thoughtful encouragement provided a foundation of stability on which this work could grow.

I am sincerely grateful to my collaborators Dr. Dengxin Dai and Dr. Shaoshuai Shi, whose expertise and openness contributed greatly to the development of the research presented in this thesis. I further wish to thank our collaborators at Toyota Motor Europe, Dr. Daniel Olmeda Reino and Dr. Rahaf Aljundi, for their insightful feedback and the stimulating exchange of ideas that broadened the impact and perspective of this work.

My appreciation also extends to former collaborators Professor Christian Theobalt, Dr. Soshi Shimada, and Dr. Vladislav Golyanik, whose involvement in earlier projects helped lay conceptual and technical foundations that continued to resonate in later stages of my research. I would also like to acknowledge Dr. Xuan Wang, Professor Yu Guo, Lichen Ma, Professor Peilin Jiang, and Professor Fei Wang from my earlier research environment, whose discussions, experiments and companionship accompanied the formative years of my academic development.

I am grateful to the administrative staff who ensured that the practical aspects of academic life ran smoothly. In particular, I would like to thank Connie Balzert and Katharina Sophie Wacker for their patience, efficiency and unfailing kindness. I also extend my thanks to IST for their reliable and much-appreciated IT support, which has been essential throughout the years. My sincere appreciation further goes to all members of Department D2, whose discussions, feedback and daily presence created a collegial and intellectually engaging environment.

Finally, I wish to thank my parents for their unwavering support and for providing me with the stability and comfort that allowed me to pursue my studies even in difficult circumstances. Their care and encouragement remain one of the greatest privileges of my life.

Contents

Introduction 1

Contents

Understanding the three-dimensional structure of humans and their surrounding environment from monocular observations is one of the most compelling and enduring challenges in computer vision [ZBSL17, PZDD17, MRC+17]. While significant progress has been made in 3D reconstruction using multi-view, stereo, or depth sensors, such methods often rely on expensive hardware, controlled settings, or large-scale labelled data [SF16, YLL+18, ZF18]. In contrast, monocular 3D perception—inferring spatial structure from a single RGB image or video—is more ambiguous, yet crucial for practical applications in autonomous systems, robotics, and human-centred computing, where lightweight, scalable, and robust solutions are essential [SCN05, EPF14].

This thesis investigates monocular 3D human-environment understanding with a focus on interaction, adaptability, and reconstruction. Rather than treating humans and their environment as separate entities, this thesis explores how the two are intrinsically coupled—how humans perceive, respond to, and shape their environment, and how environmental cues can in turn inform human motion understanding. The core motivation lies in leveraging this mutual relationship to extract richer 3D understanding from limited visual input, without requiring dense supervision or specialised hardware.

The thesis is structured as a progressive exploration of human-environment understanding from monocular inputs, organised into two main parts. Part I focuses on modelling human motion in context, beginning with how environmental cues—such as contact, support, or spatial constraints—can enhance monocular 3D pose estimation. It then explores how human motion itself can reveal properties of the environment, enabling the reconstruction of deformable or dynamic scenes through joint reasoning over human and environmental factors. Part II shifts the focus toward the reconstruction of the environment itself, addressing the challenges of scene perception under real-world constraints. This part begins by tackling the robustness of monocular depth estimation under domain shift, introducing a source-free, self-supervised test-time adaptation strategy. It then moves to single-view semantic 3D scene reconstruction, proposing rendering-based objectives and interaction-aware learning techniques to recover structured, semantically meaningful representations of the world from monocular images.

Across these stages, the methods in this thesis share a common emphasis on minimal supervision, physical plausibility, and unified geometric-semantic reasoning. Many of the proposed models are optimisation-based or self-supervised, trained without 3D ground truth, and designed to require low-cost sensor setups. This focus enables deployment in real-world scenarios while

maintaining strong geometric fidelity. By jointly considering the roles of the human and the environment, and by adapting to their changes over time, this thesis presents a pathway toward scalable, interaction-aware 3D understanding from monocular input.

1.1 Monocular 3D Human Motion Capture

Reconstructing human motion from monocular images or videos is a long-standing goal in computer vision [MHK06, Pop07]. It has broad applications in animation, sports analysis, augmented reality, surveillance, and human-robot interaction [SBB10, AGR+16]. Unlike multi-view capture systems [TGM+17, MSS+17] or marker-based motion capture setups [MHK06], monocular systems rely solely on a single RGB camera, making them lightweight, accessible, and deployable in unconstrained environments. However, this convenience comes at the cost of inherent ambiguity: depth is not directly observable, occlusions are frequent, and global scale and orientation are difficult to estimate accurately [MHRL17].

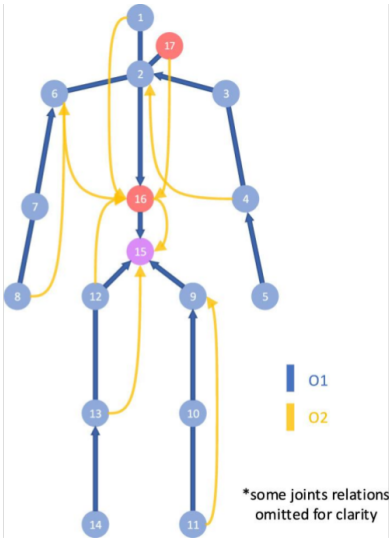
Monocular human motion capture (MoCap) aims to recover the 3D pose and motion of a person from monocular input—typically a single image or video. This task is fundamentally ill-posed: multiple 3D configurations may correspond to the same 2D projection [AB15, BKL+16]. Without multi-view geometry or depth sensors, monocular systems suffer from frequent ambiguities, especially in scale and depth ordering. In real-world scenes, additional challenges arise from motion blur, occlusion, truncation, and complex human-scene interactions [HVT+19, RBH+21].

1.1.1 Paradigms of Human Body Representations

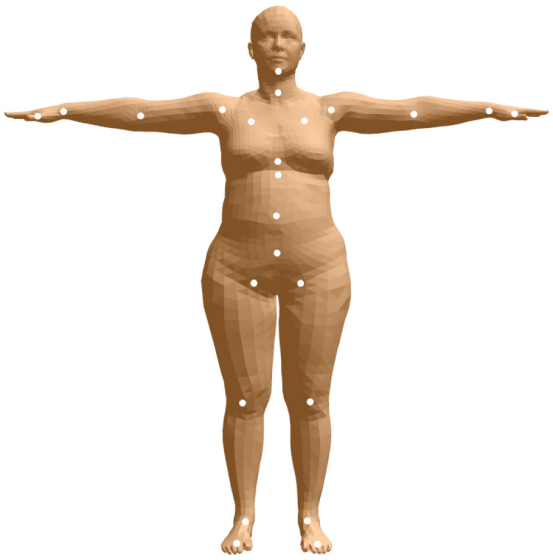
In recent years, human motion capture has advanced rapidly due to deep learning despite the challenges [ZWC+23]. Based on different forms of the human body representation, there are two dominant paradigms for human MoCap: **keypoint-based regression** [MHRL17, PZDD17, SXW+18, MRC+17, ZHS+17a] and **parametric body model fitting** [LMR+15, BKL+16, PCG+19].

Keypoint-based methods directly regress the 3D coordinates of skeletal joints (e.g., hips, knees, shoulders) from RGB inputs, as shown in Fig. 1.1a. These methods are often trained using large-scale motion capture datasets, such as Human3.6M [IPOS13] and MPI-INF-3DHP [MRC+17], to minimise joint-wise 3D loss or 2D projection error [MHRL17, PZDD17]. Their simplicity and flexibility allows for fast inference, but they often produce results that are physically implausible or unstable, due to lack of regularisation in modelling relations between joints of different human bodies.

Parametric body model-based approaches use low-dimensional representations of the human body, such as the SMPL [LMR+15, BKL+16, PCG+19] family, which encodes body shape β and pose θ into a deformable mesh model, shown in Fig. 1.1b. These models enable not only the estimation of joint locations but also full-surface geometry reconstruction. Thanks to differentiable rendering [KPBD19, KBJM18], such models can be supervised by 2D keypoints, silhouettes, and masks—without needing 3D ground truth, and can be regularised by kinematics to enforce physical plausibility. As the parametric models provide full mesh reconstruction of human bodies, they are ideal for exploring dense human-environment interactions. The methods based on parametric body models often yield more anatomically valid and temporally stable results, although they typically require optimisation and are less robust under extreme poses or out-of-distribution environments.



(a) Keypoint-based Regression



(b) Parametric Body Model (SMPL [BKL+16])

Figure 1.1: Examples of keypoint-based and parametric body model-based representations in 3D human motion capture [WTZ+21].

1.1.2 Beyond Human Body: the Role of the Environment

Although monocular MoCap methods have become more robust and accurate, most existing works still treat the human body in isolation. In contrast, real-world scenes are structured environments that strongly influence—and are influenced by—human motion. For example, body posture is shaped by contact with the ground, support surfaces, or physical constraints imposed by the surrounding scene. As such, the environment provides critical cues for disambiguating pose, resolving scale, and regularising motion [LLS+21, HVT+19].

Figure 1.2: Demonstration of human-environment interactions [HCTB19]. The human-environment contacts provide constraints to help disambiguate human localisation in 3D.

Recent research has begun to incorporate these scene-aware constraints. Some methods model ground-plane contact and object collisions [HCTB19, RGH+20, RBH+21] (see Fig. 1.2 for a demonstration provided by the PROX dataset [HCTB19]), while others directly predict contact

maps and human-object relationships [HVT+19, ZPZS21]. However, most of these efforts still assume that the environment is static, rigid, or pre-scanned—limiting their ability to generalize to dynamic or deformable settings.

In reality, many scenes are deformable (e.g., beds, couches, cushions) or evolve over time. Human motion and scene state are often mutually dependent: people adapt to their surroundings, and their interactions may alter the environment in return. Modelling such mutual influence from monocular images is still under-explored, in part due to lack of supervision and complexity of the scene dynamics.

1.1.3 Contributions

Part I of this thesis investigates monocular 3D human motion capture with a focus on environmental context and human-scene interaction, pushing beyond traditional assumptions of isolated-body estimation. The key contributions include:

- Chapter 3 presents a correction-factor-based framework for multi-person monocular 3D human pose estimation at scene scale (i.e., with absolute scale). To accurately localise the global position of the human bodies, we introduce a disentangled localisation strategy that separately models human height and body pose, leveraging scale relationships derived from 2D projection areas. This factorised formulation improves both global localisation accuracy and the quality of relative pose estimation.
- Chapter 4 proposes a scene-aware optimisation framework for monocular 3D human motion capture that enforces dense body-environment contact. To resolve the ambiguity of monocular pose estimation in cluttered scenes, the method first samples diverse candidate poses from a learned pose manifold, then refines them through contact-guided optimisation. By explicitly modelling surface-level contact between the body and scene, our approach improves both physical plausibility and absolute-scale motion accuracy, reducing artifacts such as floating and interpenetration.
- Chapter 5 introduces MoCapDeform, a novel monocular 3D human motion capture system that jointly estimates human pose and non-rigid scene deformation. Existing monocular MoCap methods often assume static environments—leading to issues like floating figures or inaccurate pose localisation when surfaces deform (e.g., sinking into a couch). In contrast, our method uses a ray-casting strategy to localise dense body–scene contact points, then performs joint optimisation over both pose parameters and dynamic scene mesh adjustments. By explicitly modelling how environments deform under human interaction, MoCapDeform achieves improved global 3D pose accuracy and dramatically reduces physical artifacts like interpenetrations, without requiring 3D annotations or depth sensors.

All proposed methods are built on monocular setups, mostly self-supervised or weakly supervised learning, using 2D labels, contact priors, or optimisation-based constraints in place of expensive 3D data. The resulting systems demonstrate robust monocular motion capture in complex, interactive, and deformable environments.

1.2 Monocular 3D Scene Reconstruction

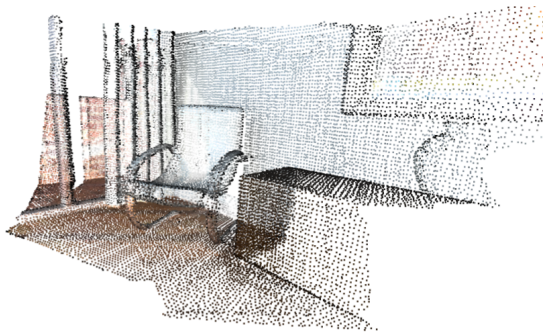
Reconstructing the geometry and semantics of a 3D scene from monocular images is a fundamental problem in computer vision [SCD+06, HZ03, XGF16]. It has a wide range of applications in

autonomous driving, robotics, AR/VR, and digital twin systems [SHKF12]. Unlike multi-view stereo or active sensing methods such as RGB-D or LiDAR scanning [ZPK18], monocular reconstruction relies solely on single RGB images or videos, making it attractive for lightweight and scalable deployment. However, this simplicity comes at a cost: the task is inherently ill-posed, as a single view does not directly encode depth or volumetric information [MWA18].

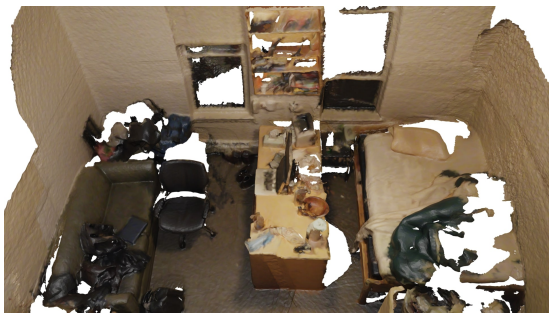
Monocular 3D scene reconstruction requires recovering spatial structure and/or semantics from limited 2D cues, under challenges such as occlusion, truncation, perspective distortion, and motion blur. In real-world scenes—especially outdoor or dynamic settings—distribution shifts (domain change of the environment) and missing 3D supervision make the task even more challenging [KPVG21, WDH+21]. To address these challenges, recent research has explored a variety of scene representations, learning paradigms, and supervision strategies. In this thesis, we focus on two key representation paradigms—depth maps and semantic occupancy grids—and introduce self-supervised methods that enable their robust estimation from monocular input without requiring 3D ground-truth annotations.

1.2.1 Paradigms of 3D Scene Representation

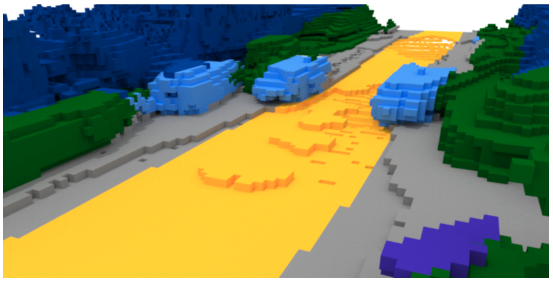
A core goal in 3D reconstruction is to recover a meaningful representation of the surrounding environment. To this end, choosing an appropriate scene representation is critical, as it defines both the reconstruction target and the inductive biases of the system. Various paradigms have been proposed, each making different trade-offs in expressiveness, efficiency, supervision requirements, and suitability for downstream tasks [SS23]. Fig. 1.3 demonstrates examples of the main families of scene representations.



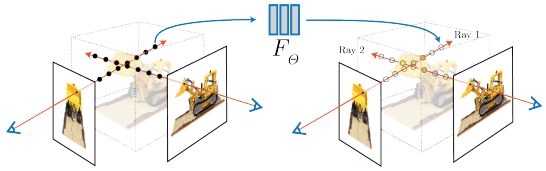
(a) Point cloud [ZPK18]



(b) Mesh [DCS+17]



(c) Voxel [HWM+24]



(d) NeRF [MST+21]

Figure 1.3: Examples of scene representations.

Point clouds represent scenes as unordered sets of 3D points—typically acquired via LiDAR or RGB-D sensors [RC11]. A structured variant is the **depth map**, which can be viewed as a

dense 2.5D point cloud aligned to the image plane, and can be converted to 3D point clouds with back-projection using camera parameters. Depth maps preserve pixel-level detail and are amenable to self-supervised learning using geometric and photometric consistency in monocular sequences [GMAFB19, LHKS19]. They are widely used as a fundamental intermediate geometric representation for monocular 3D tasks [MWA18].

Meshes represent 3D surfaces using a collection of connected vertices and faces—typically triangles—forming a polygonal approximation of object boundaries [HZ03]. This format enables compact, visually accurate surface modelling and supports applications like rendering, animation, and simulation. However, in the context of monocular 3D reconstruction, meshes pose significant challenges: scene-level meshes often have complex, scene-specific topology that is hard to predict and structure. As a result, meshes are rarely used as direct outputs of neural networks in monocular settings. Instead, they are typically produced via post-processing steps applied to intermediate representations, such as voxel grids, signed distance fields, or point clouds [SF16, NIH+11].

Voxel representations discretise space into regular volumetric grids, storing occupancy values per cell [WSK+15]. Extending this concept, **semantic occupancy grids** assign semantic labels to each voxel, enabling structured comprehension of scene layout and object presence. Recent work shows that such grids can be predicted monocularly using differentiable rendering and image-level supervision (e.g., 2D semantic maps, depth maps) without dense voxel annotations [WYRC23, PLZ+24, HZZ+24a].

Implicit neural field (e.g., NeRF [MST+21]) is a new type of scene representation proposed in recent years and quickly gains a lot of attention. Instead of explicitly reconstruct the scene geometry, it represent geometry or appearance as continuous neural functions, enabling detailed reconstructions which can later be extracted and converted to different explicit representations. The differentiable rendering capabilities make it well-suited for self-supervised learning, facilitating monocular occupancy pipelines that enforce 2D supervision via rendered projections into images.

1.2.2 3D Reconstruction with Self-supervision

Reconstructing 3D scenes from visual data has traditionally relied on dense 3D supervision—such as ground-truth depth maps, LiDAR point clouds, or multi-view stereo reconstructions—which are expensive to obtain and difficult to scale across diverse environments [SCD+06, NIH+11]. Collecting accurate 3D labels requires specialised hardware, careful calibration, and labour-intensive annotation pipelines, making such methods impractical for large-scale deployment. In contrast, self-supervised learning seeks to infer 3D structure from raw monocular inputs by exploiting geometric priors and naturally occurring signals—such as view consistency or semantic alignment—without relying on explicit 3D annotations. This paradigm has become increasingly important for building flexible, scalable 3D perception systems based solely on widely available monocular imagery.

A well-established line of research focuses on self-supervised monocular depth estimation. Here, depth is inferred by reconstructing target views using estimated depth and pose, minimising photometric reprojection error between real and synthesised images [GMAFB19, BLW+19]. To improve geometric accuracy, auxiliary losses such as smoothness, depth gradients, and temporal consistency are often employed [WMAP+21]. These methods have enabled learning dense, metric depth maps without any 3D supervision and have been successfully applied to outdoor driving datasets like KITTI [GLSU13]. However, such systems often suffer from poor generalisation when deployed in new domains, motivating recent research on test-time adaptation and domain-agnostic learning.

Another branch of research aims at full 3D reconstruction, e.g. semantic occupancy prediction from monocular images, which goes beyond surface geometry to infer volumetric structure and semantics. These methods typically represent the scene as a 3D voxel grid and supervise the learning through 2D rendering losses, aligning the projections of predicted 3D volumes with semantic segmentation or depth maps [ZZD23, PLZ+24]. By integrating rendering-based supervision, these methods bypass the need for voxel-level ground truth, making them suitable for scalable training in real-world scenarios. More recently, differentiable scene representations such as NeRF [MST+21] and 3D Gaussian Splatting [KKLD23] have enabled efficient self-supervised training pipelines that link 2D image supervision with 3D geometry in a differentiable way. While many of these frameworks focus on novel view synthesis, they also offer a pathway for unsupervised scene-level reasoning.

Despite their promise, self-supervised 3D reconstruction methods still face several challenges. First, monocular systems inherently lack information about occluded or distant regions, limiting the completeness of reconstruction. Second, domain shifts across lighting conditions, camera parameters, or environments can significantly degrade performance, particularly for outdoor scenes. Finally, a key limitation lies in the absence of integrated modelling between geometry and semantics—despite their natural complementarity, few existing methods explicitly leverage interactions between the two to improve overall 3D understanding.

These limitations motivate the development of more robust, adaptive, and semantically structured frameworks for monocular 3D scene reconstruction—without relying on dense labels or specialised hardware.

1.2.3 Contributions

Part II of this thesis investigates monocular 3D scene reconstruction, focusing on practical and scalable representations that capture geometry and semantics from single-view images. The key contributions include:

- Chapter 6 introduces a source-free, test-time adaptation framework for monocular depth estimation. Existing monocular depth models often suffer severe performance degradation when deployed across domains with different appearance statistics (e.g., weather, lighting, location). To address this, we propose a self-supervised optimisation pipeline that adapts model predictions to target domains using only photometric and geometric consistency at test time—without requiring source data, annotations, or retraining. The method shows strong improvements under real-world distribution shifts, especially in outdoor driving scenes, and requires no architectural changes or domain-specific tuning.

- Chapter 7 proposes a semantic 3D occupancy reconstruction framework from a single monocular image, using only 2D supervision. We build a differentiable volumetric rendering system inspired by neural implicit fields, aligning voxel-level semantic predictions with 2D semantic maps through rendering losses. To bridge the gap between geometry and semantics, a multi-task interaction module is designed to coordinate learning across tasks and improve 3D consistency. This system achieves accurate semantic voxel reconstruction without any 3D occupancy ground truth, relying solely on monocular RGB and 2D annotations for supervision. The resulting representation supports efficient single-frame inference and enables rich semantic understanding of the 3D scene.

Both methods operate under self-supervised or weakly supervised regimes and require no access to ground-truth 3D geometry, depth, or occupancy labels. Together, they demonstrate

how scalable monocular pipelines can recover meaningful 3D structure and semantics, even in the face of domain shift and supervision sparsity.

1.3 Outline

This thesis is divided into 8 chapters, organised in 2 parts:

Chapter 2, Related Work:
We review previous work directly related to the topics discussed in this thesis. These include 3D human pose estimation and motion capture, human-scene interaction and contact reasoning, monocular depth estimation and domain adaptation, and 3D occupancy prediction.

Part I, Capturing 3D Human Motion from Scenes

Chapter 3, Multi-person Pose Estimation in Scene Scales:
We present a correction factor based framework for scene-scale multi-person monocular 3D human pose estimation, enabling recovery of absolute-scale global positions. Central to the method is a disentangled localisation scheme that decouples human height from body pose, guided by geometric cues inferred from 2D projection areas. This factorised design enhances both the accuracy of global localisation and the consistency of relative pose estimation across individuals.

This chapter corresponds to the CVIU publication [GML+21] with the title *Monocular 3D Multi-person Pose Estimation via Predicting Factorized Correction Factors*. As the third author, Zhi Li contributed to the development of the factorised correction–factor idea, drawing from her previous work [LWWJ19] and master thesis, and was involved in key technical discussions and implementation aspects of the paper.

Chapter 4, Human Motion Capture from Scene Contact Guidance:
We introduce a scene aware optimisation method for monocular 3D human motion capture, which leverages dense body-environment contact to resolve pose ambiguities in cluttered scenes. The approach begins by drawing diverse pose candidates from a learned manifold and subsequently refines them through contact-driven constraints. By explicitly incorporating surface-level interactions with the environment, the method enhances physical realism and improves global accuracy, effectively mitigating issues such as floating bodies and interpenetrations.

This chapter corresponds to the ECCV publication [SGL+22] with the title *HULC: 3D Human Motion Capture with Pose Manifold Sampling and Dense Contact Guidance*. As the third author, Zhi Li contributed to the development of the dense-contact generation components, drawing on work conducted in parallel with the first-author project on deformable-environment motion capture [LSS+22] (Chapter 5). Zhi Li also participated in technical discussions that informed the shared contact modules and implemented key parts of the GTA-IM annotation pipeline, including tracking and SMPL parameter optimisation.

Chapter 5, Human Motion Capture with Scene Deformation Recovery:
We present MoCapDeform, a monocular 3D human motion capture system that simultaneously recovers human pose and non-rigid scene deformations. Unlike prior methods that assume a static environment—often resulting in floating artifacts or mislocalised poses

when soft surfaces are involved—this approach captures the mutual influence between the human and their surroundings. A ray-casting mechanism is used to infer dense contact points, which guide a joint optimisation over both body pose and scene mesh adjustments. By accounting for environment deformation induced by human interaction, MoCapDeform significantly enhances pose accuracy and reduces physical inconsistencies, all without relying on 3D supervision or depth inputs.

This chapter corresponds to the first-author 3DV publication [LSS+22] with the title *MocapDeform: Monocular 3D Human Motion Capture in Deformable Scenes*. The paper has won the *Best Student Paper Award* in 3DV 2022.

Part II, Reconstructing and Understanding 3D Scenes

Chapter 6, Test-time Adaptation for Monocular Depth Estimation:
We present a test-time adaptation strategy for monocular depth estimation under domain shift, designed to operate in a source-free setting. Traditional depth models trained on one domain often degrade significantly when applied to visually different environments due to changes in lighting, weather, or scene layout. To overcome this, a self-supervised optimisation framework is introduced that refines model predictions directly at inference time using only geometric and photometric consistency, without access to source domain data or annotations. This plug-and-play approach delivers substantial improvements in cross-domain scenarios—particularly in challenging outdoor driving scenes—while requiring no changes to network architecture or offline retraining.

This chapter corresponds to the first-author ICRA publication [LSSD23] with the title *Test-time Domain Adaptation for Monocular Depth Estimation*.

Chapter 7, 3D Occupancy Prediction via Multi-task Distillation:
We introduce a framework for reconstructing semantic 3D occupancy from a single monocular image, supervised only with 2D annotations. Drawing inspiration from implicit neural representations, the method leverages a differentiable volumetric rendering pipeline to align voxel-wise semantic predictions with image-level semantic maps via rendering-based losses. To enhance coherence between geometric structure and semantic content, a dedicated multi-task interaction module is incorporated, promoting joint learning and improving 3D consistency. Without requiring any ground-truth 3D occupancy labels, the system achieves high-quality reconstructions and supports efficient, single-frame inference with rich semantic awareness of the scene.

This chapter corresponds to the first-author GCPR publication [LARS25] with the title *MT-Occ: Single-view 3D Occupancy Prediction via Multi-task Distillation*.

Chapter 8, Conclusion:
We conclude this thesis by summarising our key findings and discussing promising directions of future work.

1.4 Publications

The content of this thesis has previously appeared in the following publications, ordered as outlined above:

[GML+21] Yu Guo, Lichen Ma, Zhi Li, Xuan Wang and Fei Wang. Monocular 3D multi-person pose estimation via predicting factorized correction factors. *Computer Vision and Image Understanding (CVIU)*, 213 (2021): 103278.

[**SGL+22**] Soshi Shimada, Vladislav Golyanik, Zhi Li, Patrick Pérez, Weipeng Xu, Christian Theobalt. Hulc: 3d human motion capture with pose manifold sampling and dense contact guidance. *European Conference on Computer Vision (ECCV)*, 2022.

[**LSS+22**] Zhi Li, Soshi Shimada, Bernt Schiele, Christian Theobalt, Vladislav Golyanik. Mocapdeform: monocular 3d human motion capture in deformable scenes. *International Conference on 3D Vision (3DV)*, 2022.

[**LSSD23**] Zhi Li, Shaoshuai Shi, Bernt Schiele, Dengxin Dai. Test-time domain adaptation for monocular depth estimation. *International Conference on Robotics and Automation (ICRA)*, 2023.

[**LARS25**] Zhi Li, Rahaf Aljundi, Daniel Olmeda Reino, Bernt Schiele. MT-occ: single-view 3D occupancy prediction via multi-task distillation. *German conference on pattern recognition (GCPR)*, 2025.

Related Work 2

Contents

In this chapter, we review literature on relevant prior work across four key domains that support the core directions of this thesis. First, we discuss monocular 3D human pose and motion estimation, focusing on absolute-scale and parametric model based methods. Next, we examine approaches for integrating scene context and contact reasoning into motion capture. The third section covers monocular depth estimation and domain adaptation, particularly under source-free and test-time settings. Finally, we survey 3D scene representation paradigms and recent progress in learning-based occupancy prediction from monocular views. The chapter concludes by summarising how this thesis positions itself with respect to these lines of work.

2.1 Monocular 3D Human Pose and Motion Estimation

Monocular 3D human pose estimation aims to infer the 3D body configuration of one or more individuals from a single RGB image or video stream. This task is fundamentally ill-posed due to the inherent ambiguities of monocular vision, such as the absence of direct depth cues, scale ambiguity, occlusion, and truncation. Over the past decade, the field has progressed through several paradigms—from 2D-to-3D lifting to parametric mesh regression, from isolated single-person estimates to dynamic multi-person motion capture in cluttered, realistic scenes.

Keypoint-based Regression. A foundational line of work formulates 3D human pose estimation as a regression problem over anatomical keypoints [SBIK16, ZWC+23]. These methods typically predict root-relative joint coordinates from 2D keypoints detected in the image. Pioneering efforts in this area include simple yet effective architectures that regress 3D pose from 2D joint heatmaps [PZDD17, PZZD18, PZD18], or directly from image features [MHRL17, CR17, ZHS+17b]. To address the spatial uncertainty in 2D-to-3D lifting, volumetric representations such as voxel heatmaps have been proposed [PZDD17, MCL19], allowing networks to better capture depth cues. Several methods also explore intermediate feature encodings, such as integral pose regression [SSLW17, SXW+18] or graph convolution over skeletal structures [BCW+20]. Later improvements include camera-aware regression [WR19, WLR22] and depth estimation-enhanced learning [PZD18, VL19]. Despite strong empirical performance, keypoint-only approaches often lack structural coherence and struggle to produce physically plausible body configurations, especially in occluded or cluttered environments [SBIK16, ZWC+23].

Parametric Body Models. To improve anatomical realism and enable dense body surface reconstruction, parametric body models like the SMPL family [LMR+15, BKL+16, PCG+19] have become standard. These models represent the human body as a mesh deformable via pose and shape parameters, enabling the use of optimisation [BBLR15, BKL+16] or deep regression pipelines [KBJM18, KPBD19] to estimate 3D human mesh from monocular images. Subsequent

works improve robustness and expressiveness through iterative refinement [KPJD21], adversarial training [WMM+21], or hybrid analytical-neural models [LLS+21]. Mesh-based methods also support integration of body part priors [JSS18], learned attention [ZPZS21], or volumetric alignment [GLK+20]. Temporal extensions [KAB20] further improve consistency over video sequences. However, most regress pose relative to a canonical frame, leaving global positioning under-constrained.

Scene-scale Pose Estimation. Estimating 3D pose in global coordinates introduces additional challenges: resolving absolute scale, understanding camera placement, and grounding pose in the scene [SYW+23]. Prior works attempt to infer scene-scale pose via geometric consistency [RSF18], vanishing point estimation [DMK+18, DGM+19], or scene affordances [HVT+19]. Recent methods localise humans relative to ground planes [SGXT20, SGX+21, RGH+20] or leverage depth maps or known camera geometry [DMK+18, LWZ+21]. Nevertheless, many remain limited to single-person or rely on auxiliary inputs. In contrast, our approach introduces a correction-factor-based formulation to disentangle human height and pose in multi-person scenes using scale-inferred cues from 2D projections.

Multi-person 3D Pose Estimation. Multi-person scenarios present additional complexity, such as joint occlusion, overlapping limbs, and ambiguous identities [FXTL17]. Early efforts extend 2D pose detectors with 3D lifting per person [RWS17, RWS19, MCL19], while others adopt holistic frameworks that jointly infer multiple 3D poses [ZMS18, VL19]. Top-down pipelines first detect bounding boxes then estimate pose [RWS17, RWS19, MCL19], whereas bottom-up methods directly detect part-level keypoints and group them into skeletons [ZOL+20, HLX+23, CHS+19]. However, many methods still operate in root-relative coordinates, ignoring global alignment. In our work, we extend global localisation to multi-person scenes via a scale-disentangled representation that separates height estimation from pose regression, improving both scene grounding and inter-person consistency.

Video-based Human Motion Capture. Monocular video-based human motion capture (MoCap) extends the goal of 3D pose estimation from isolated frames to temporally consistent motion recovery across sequences. While 3D pose estimation typically refers to inferring the skeletal joint positions of a person in a single image [MHRL17, PZDD17, ZBSL17], motion capture requires recovering temporally smooth, physically plausible 3D trajectories of body motion over time, often including full-body mesh representations [KBJM18, KPBD19, KAB20]. The transition from frame-wise pose estimation to full MoCap introduces several new challenges: maintaining temporal consistency, resolving motion blur and occlusions, and enforcing biomechanical realism. Early approaches relied on recurrent neural networks (RNNs) to incorporate temporal priors and smooth transitions [HL18], while others leveraged optical flow to enforce inter-frame correspondence [ADZ19]. Temporal convolutional networks (TCNs) were introduced for more stable long-term modelling with fewer parameters [PFGA19]. [KAB20] proposed a seminal adversarial MoCap framework that combines a temporal encoder with a discriminator trained on real MoCap data to generate realistic motion sequences from monocular input. Further improvements have leveraged transformer-based architectures for sequence modelling, such as action-conditioned transformer VAE [PBV21], and multi-frame SMPL regression models like TCMR [CML21], which refine temporal priors with 3D supervision. Some methods integrate temporal mesh refinement or temporal self-supervision to reduce jitter and improve physical consistency [RPKM21]. However, most of these models still operate under the assumption of a static scene, ignoring how motion is influenced by body–environment interaction. Recent works have begun incorporating physical contact and environmental awareness into the motion capture pipeline. For instance, [RBH+21] introduced a learned motion prior that enforces plausible contact and surface interaction. [CPB+20] proposed a part-guided attention mechanism to better

align body motion with surrounding objects. This thesis builds upon these developments by introducing scene-aware and deformation-sensitive monocular MoCap frameworks. Unlike prior work that treats the environment as static or ignores it altogether, this thesis explicitly models dense contact points and scene deformation, offering improved realism and physical fidelity in cluttered or dynamic environments—all from monocular RGB input.

2.2 Human-Scene Interaction and Contact Reasoning

The estimation of 3D human pose and motion in real-world environments often requires accounting for the rich and complex interactions between the human body and surrounding objects or surfaces. Unlike traditional monocular MoCap methods that treat the human body in isolation [CR17, MHRL17, TRA17, MN17, FXW+18, DGM+19, TKS+16, MRC+17, RSF18, MSS+17, PZDD17, HXM+19], recent efforts incorporate physical contact, environmental constraints, and scene context to improve realism, stability, and scale-aware reconstruction [KBJM18, PZZD18, PCG+19, KAB20, ZHH+21, MCL19, SAA+20, MSM+20, SGXT20, SGX+21, RGH+20, RBH+21]. This section reviews relevant literature in contact-aware motion capture, dense interaction modelling, scene-aware pose estimation, and joint reasoning over humans and environments.

Scene-Aware MoCap and Human-Environment Constraints. To resolve inherent ambiguities in monocular MoCap and ensure physical plausibility, an emerging direction incorporates environmental context and scene constraints. Several works enforce contact with the ground plane, either by detecting foot-floor contact or by constraining body volume to avoid penetration [ZMS18, SAA+20, RGH+20, SGXT20, RBH+21]. Other methods utilise holistic scene representations constructed by arranging object templates or recovering scene layouts [MGC+19, ZPJ+20, WY21]. These enable coarse reasoning about human position in the global space, yet rely on object category assumptions or lack surface-level contact precision. More detailed scene-aware methods make use of pre-scanned 3D meshes of the environment [HCTB19, ZZB+21, WCR+22, CGM+20]. PROX [HCTB19] introduces a signed distance field (SDF) representation to reason about proximity and contact between human and scene, while LEMO [ZZB+21] learns motion priors from RGB-D inputs to improve realism and robustness under occlusion. Although these methods facilitate better localisation, they depend heavily on high-quality scene scans and struggle with dynamic or deformable environments.

Contact-Driven Human–Scene Interaction. Contact reasoning lies at the heart of reconstructing physically plausible 3D human motion, especially to resolve depth ambiguities inherent in monocular capture. Early works regress sparse contacts—often on joints or kinematic skeletons—or apply hand-crafted heuristics to detect ground or object contacts [LSC+19, SGXT20, SGX+21, RGH+20, RBH+21, ZYC+20, ZZB+21]. [ZZB+21] further demonstrated sparse contact detection via RGB-D inputs and segmentation guidance. In contrast, recent endeavours aim to predict contacts on dense human mesh surfaces: [HGT+21a] propose a semantically-informed dense-contact dataset for body–scene placement, while [MOT+21] introduce a self-contact loss to enforce plausible, non-penetrating body configurations. However, these approaches either address only self-contact or treat the scene context in oversimplified ways without modelling the environment's contribution to contact. More advanced methods such as [RBH+21, BTT+20] incorporate ground contact into the motion prior, enabling contact-aware pose refinement from 2D/3D observations and RGB(-D) inputs. [SGXT20, SGX+21] use physics-based trajectory optimisation to enforce plausible human-ground dynamics via latent contact signals. Outside whole-body interactions, [BTT+20] provides a rich hand–object contact dataset that supports learning contact correspondences between hand surface and object geometry via RGB-D input. Additionally, POSA [HGT+21b] learns body–scene vertex-wise contact probabilities from SMPL-

X and scene geometry to improve placement realism and affordance reasoning. Nevertheless, none of these methods jointly estimate dense contacts for both the human body and its surrounding environment solely from monocular RGB and static scene geometry. This thesis contributes to this line of work by leveraging both dense scene geometry and learned priors to infer contact-aware human motion from minimal input.

Scene Deformation and Global Human Motion Capture. Most existing scene-aware MoCap approaches assume a rigid, static environment [HCTB19, ZZB+21, HCTB19, RGH+20, RBH+21], and constrain human motion by registering against fixed mesh surfaces. While such rigid-scene constraints facilitate global localisation and collision avoidance, they fall short in dynamic or deformable environments commonly found in the real world, where objects may bend, compress, or shift in response to human interaction. Existing work typically do not explicitly model scene deformation. Notably, PROX [HCTB19] and LEMO [ZZB+21] assume a high-fidelity SDF representation of the scene, which limits their applicability to pre-scanned, unchanging environments. Although volumetric representations such as voxelised scenes [WCR+22] or point clouds [WY21] offer greater flexibility, they rarely support continuous deformation modelling. The ability to reason about scene deformations remains largely unexplored in monocular MoCap literature. Yet, such modelling is crucial in cases where humans interact with soft furniture or objects that respond to contact forces. This thesis addresses this gap by integrating deformation-aware reasoning into the MoCap pipeline, facilitating accurate global human motion reconstruction even under non-rigid environmental changes.

2.3 Monocular Depth Estimation and Domain Adaptation

The task of predicting dense 3D geometry from single-view RGB input is a core component in many 3D perception systems. In particular, monocular depth estimation plays a vital role in enabling scene-level understanding from video or image streams without the need for expensive active sensors such as LiDAR. However, learning-based monocular depth estimation remains highly sensitive to domain shifts: models trained in one domain (e.g., clear weather or synthetic environments) often degrade significantly when deployed in another (e.g., rainy or night-time driving scenes). These challenges necessitate robust learning strategies that generalise across domains. We review representative works in both supervised and self-supervised monocular depth estimation, followed by domain adaptation techniques—particularly those under test-time and continuous adaptation settings—that aim to mitigate performance drops due to distribution shift.

Monocular Depth Estimation and Self-supervision. Early efforts in supervised depth estimation directly regress dense depth maps from RGB images using paired ground-truth LiDAR as supervision. Eigen et al. [EPF14] propose a multi-scale deep network to predict depth maps with fine-grained resolution, pioneering deep learning for this task. Subsequent works focus on architectural innovations and auxiliary signals to enhance accuracy. For instance, Lee et al. [LHKS19] propose Big-to-Small networks for learning better global context, while Aich et al. [AVIL21] design bi-directional feature fusion modules. Further advancements such as Patch-wise Depth Completion [LLK+21], SiDERT [SCC+22], and BinsFormer [LWLJ22] aim to improve spatial detail recovery and scale prediction. Other works integrate cross-modal constraints to guide learning, using semantic segmentation and surface normals [YLSY19, ZCX+19], or incorporate temporal consistency across video frames [ZSL+19]. Despite their high performance, supervised methods are constrained by their reliance on dense ground-truth depth, which is expensive to collect and difficult to scale. **Self-supervised approaches** remove the need for ground-truth depth by treating depth estimation as a view synthesis problem. Zhou et al. [ZBSL17] and

Godard et al. [GMAFB19] train depth and pose networks jointly by minimizing the photometric reconstruction loss between the reference and synthesized views. Various enhancements have been proposed to address limitations of the basic framework. Chen et al. [CSS19] and Luo et al. [LHS+20] improve geometric consistency through auxiliary optical flow estimation. Guizilini et al. [GAP+20] incorporate weak supervision from vehicle velocity to recover metric depth. Watson et al. [WMAP+21] exploit temporal cost volumes to improve dynamic scene handling. Despite their scalability, self-supervised methods suffer from scale ambiguity—inferred depths are up to an unknown scale and require alignment using ground-truth depth during inference unless additional scale recovery methods are applied, e.g., [XZH+20].

Domain Adaptation for Monocular Depth Estimation. Domain adaptation (DA) techniques aim to bridge the distribution gap between a labelled source domain and an unlabelled target domain [PTKY10, PGLC15]. Two common strategies are feature alignment [LCWJ15, GL15, THS+18] and self-training via pseudo labels [ZYKW18, LLDG19, WDH+21, HDVG22]. Input-level style transfer is also used, such as CycleGAN-based CyCADA [HTP+18] or FDA [YS20]. These methods, while effective for classification and segmentation, often struggle to scale to dense prediction tasks like depth estimation. Notably, Tonioni et al. [TPMDS19] introduce an online adaptation framework specifically for depth, while Kuznietsov et al. [KPVG21] explore adaptation under continual domain shift.

Test-Time Adaptation. Test-time domain adaptation (TTDA) restricts access to source data during adaptation [KVB+20, YWvdW+21]. This scenario is particularly relevant for on-device or privacy-sensitive deployment. Feature alignment becomes challenging, so methods instead focus on updating the target model using entropy minimization [WSL+20, LHF20] or batch statistics [KECK21, ZL21, HUC+21, YLZ21]. Others use generative models to align feature distributions [KSN21, LJC+20, YYYH21]. However, these methods are mostly designed for classification or segmentation, with limited extensions to depth.

Continuous Domain Adaptation. In contrast to fixed target domain setups, continuous domain adaptation handles gradually evolving domains over time [WFVGD22]. Approaches like incremental adversarial adaptation [WBP18, BTHD18] have been proposed for robotics and semantic tasks. Wang et al. [WFVGD22] address the combination of source-free, continual test-time adaptation, but focus solely on image classification. For depth estimation, Kuznietsov et al. [KPVG21] propose COMODA, a continual depth adaptation method that assumes access to ground-truth velocities in the target domain—a non-trivial requirement for practical deployment.

Overall, robust monocular depth estimation under domain shift remains an open challenge. Self-supervised frameworks offer a scalable solution but require scale recovery and may not generalise across diverse appearances. Domain adaptation—particularly test-time and source-free variants—provides promising directions, though applications to dense prediction tasks such as depth are still underexplored. The gap motivates the development of lightweight, source-free adaptation pipelines that leverage geometric constraints without requiring access to target ground-truth or source-domain data.

2.4 3D Scene Representation and Occupancy Prediction

As 3D perception becomes central to many vision applications such as autonomous driving, robotics, and AR/VR, the challenge of learning accurate, efficient, and semantically meaningful 3D scene representations has drawn increasing attention. Among these, voxel-based semantic scene completion (SSC) and 3D occupancy prediction are prominent tasks aiming to jointly estimate geometry and semantics in volumetric space. This section reviews representative works along several axes: 3D representation types, semantic scene completion, self-supervised 3D

occupancy learning, neural field-based methods, and the role of multi-task supervision.

3D Scene Representations. Existing 3D scene understanding methods rely on various representations, each balancing trade-offs between fidelity, efficiency, and scalability. Traditional grid-based volumetric representations [SYZ+17, LLG+19] allow dense geometry and semantics estimation but suffer from cubic memory complexity. Point-based methods [QSMG17, QYSG17] enable sparse and efficient computation but are less suited to regular voxel-level predictions. Mesh-based approaches [KTEM18, PKS+19] provide high fidelity but are hard to use for dense semantic tasks. Implicit representations, especially Neural Radiance Fields (NeRFs) [MST+21], have recently emerged as a powerful alternative that encode scenes in continuous volumetric functions, enabling high-resolution view synthesis and geometry estimation. These representations are particularly attractive in self-supervised settings, where 2D supervision can be lifted into 3D via differentiable rendering.

Semantic Scene Completion and 3D Occupancy Prediction. SSC refers to predicting both occupancy and semantic labels in a volumetric grid, typically aligned to a fixed coordinate system. Early SSC methods fused depth or RGB-D data into voxels and predicted semantics using 3D convolutions [SYZ+17, LLY+19, LLG+19, LHW+20]. Variants like S3CNet [CAR+21], LMSCNet [RdCVB20], and SSCBench [LLL+24] expand upon this framework by introducing sparse convolutions [YGL+21], anisotropic kernels [LHW+20], or RGB-guided feature fusion [LHZ+18, WTNT20]. Recently, occupancy prediction using vision-only inputs, termed "3D occupancy prediction," has gained traction in autonomous driving [TSW+23, WZX+23], where LiDAR may not always be available. Multi-view models [PLZ+24, WZZ+23, HZZ+24b, HTZ+25] leverage camera rigs to generate BEV (Bird's Eye View) features and predict voxel occupancies. In contrast, monocular methods such as MonoScene [CDC22] and S4C [HWM+24] tackle the harder problem of inferring 3D structure from a single frame, often incorporating geometric priors or temporal smoothness.

Self-supervised Occupancy Prediction. A notable development in 3D occupancy learning is the use of self-supervised or weakly-supervised strategies. Rather than relying on voxel-level ground truth, methods like OccNeRF [ZYW+23], SelfOcc [HZZ+24a], and RenderOcc [PLZ+24] reconstruct 3D geometry by supervising view-consistent predictions using 2D semantic labels and posed images. This paradigm is typically implemented via neural volume rendering, inspired by NeRFs [MST+21], enabling gradients to propagate from 2D loss functions into 3D representations. Some methods further distill semantic information from multiple views into 3D features [HZZ+24a, ZYW+23], while others explore Gaussian splatting-based representations [GLX+24, CZB+25, BGR25, ZWW+25, KKLD23], enabling efficient rasterization-based rendering. While many self-supervised occupancy methods focus on multi-view settings, S4C [HWM+24] proposes a monocular variant that trains with multi-view supervision but infers from a single frame at test time, showing promise for practical deployment.

Leveraging Depth and Semantic Priors. Depth estimation and semantic segmentation are powerful intermediate tasks that can guide 3D reasoning. Several works show that monocular or stereo depth predictions can improve NeRF training by stabilising volume sampling or regularising geometry [YPN+22, YHL+23, PHT23, WLZ+24, CHL+24]. Semantic supervision, through segmentation or panoptic labels, can be distilled into volumetric fields [FZC+22, ZLLD21, KGY+22] or used to separate geometry and semantics during rendering [PGJ+23, LFS+24]. Recent works [FYJ+22, KKG+23] also incorporate language-driven priors to enable generalizable scene understanding. These priors are often complementary, as semantics help resolve ambiguities in structure (e.g., wall vs. floor), while depth guides spatial reasoning. Our work exploits this synergy by jointly training with multiple 2D tasks, distilling their features into the 3D representation space.

Multi-task Learning for 2D-to-3D Prediction. Multi-task learning provides a principled framework for sharing features across related tasks such as depth estimation, segmentation, and motion prediction. Works like PAD-Net [XOWS18], MTI-Net [VVGG+20], MultiMAE [BMAZ22], and MTFormer [XZV+22] demonstrate the benefits of task interaction for improving accuracy and robustness. Techniques such as task affinity modelling [LLJ+21], decoupled heads [BZSS22], and cross-modal transfer [LVdC23] allow networks to adaptively prioritise tasks. While these works primarily focus on 2D outputs, our work adapts the same principles to guide 3D occupancy learning, using multi-task distillation to bridge the gap between 2D supervision and 3D reconstruction. In particular, we show that learned representations from depth and semantics can be projected into the 3D domain to improve occupancy inference under sparse or ambiguous inputs.

Together, these lines of research inform our proposed approach, which unifies monocular 2D perception and 3D scene understanding through efficient representation learning, self-supervised training, and cross-task distillation.

2.5 Summary and Positioning of this Thesis

This thesis addresses the long-standing challenge of monocular 3D human motion capture and scene understanding, with a particular focus on modelling human-scene interactions from sparse, single-view inputs. While monocular MoCap has seen significant progress in recent years—with methods leveraging temporal cues, parametric body models, and self-supervised learning—most existing approaches operate under restrictive assumptions, such as static backgrounds or ground plane contact only. They often fail to handle complex interactions between humans and their surrounding 3D environments, resulting in physically implausible poses, depth ambiguity, or unrealistic motion dynamics.

To bridge this gap, this thesis proposes a series of learning-based frameworks that extend monocular motion capture to physically-aware, contact-sensitive, and geometry-aware settings. We introduce methods that jointly estimate human motion, dense contact patterns, and environmental geometry, enabling more plausible reconstruction of interactive human behaviour in diverse scenes. By leveraging self-supervised strategies, our approaches significantly reduce the reliance on dense labels for depth estimation when adapting to diversely changing environments. Furthermore, we tackle the complementary problem of single-view 3D occupancy prediction by distilling rich 2D priors—such as semantics and depth—into 3D volumetric representations, demonstrating that sparse monocular cues can support high-fidelity scene reconstruction and semantic reasoning.

Positioned at the intersection of 3D human modelling, scene understanding, and self-supervised learning, this thesis contributes to a new generation of monocular perception systems that move beyond isolated human pose estimation toward holistic and interactive 3D scene reconstruction. Compared to prior works, our methods handle more general environments, require less supervision, and offer unified solutions for contact reasoning, depth estimation, and scene completion—laying the foundation for practical deployment in robotics, AR/VR, and embodied AI.

I

Capturing 3D Human Motion from Scenes

3 Multi-person Pose Estimation in Scene Scales

Contents

Despite the great achievement of 3D human pose estimation, recovering the 3D poses of multiple persons in absolute scene scales with localisation from a single image is still a challenging problem. In this chapter, we focus on one specific problem in 3D multi-person pose estimation (3D-MPPE): estimating the absolute 3D human poses. We proposed a pipeline consists of human detection, absolute 3D human root localization, and root-relative 3D single-person pose estimation modules. For the absolute 3D human root localization task, we propose a decoupling dual-branch structure to reconstruct the height of the human body, and further output the depth and localization of the 3D human root in the camera coordinate system. Furthermore, a data augmentation strategy is presented to tackle occlusions, such that our model can effectively estimate the root localization with the incomplete bounding boxes. For the 3D human relative pose estimation task, we use the attention mechanism to capture the correlation between human joint coordinates and further improve the accuracy of relative pose estimation. Finally, we merge the absolute depth of human and the relative 3D pose to output the absolute 3D human pose.

This chapter is based on [GML+21]. As the third author, Zhi Li contributed to the main idea of using factorised correction factor to approximate absolute human sizes as this work represents a significant extension of Zhi Li's previous paper [LWWJ19] and master thesis. In addition, Zhi Li contributed to detailed technical discussions and implementations, and was involved in paper writing and visualisations.

3.1 Introduction

Estimating the 3D poses of multiple persons from a monocular image is an important computer vision task with extensive applications, such as action recognition, motion capture, etc. With the emergence of deep learning techniques, great progress has been achieved on this topic in recent years.

Most existing methods focus on the 3D single-person pose estimation (3D-SPPE) case

[FXW+18, HGT17, JY17, MRC+17, LWWJ19], in which the root-relative pose (i.e. relative positions of joints w.r.t a pre-defined root joint of a human body) is predicted from a cropped image yielded by detection. Only a few approaches [MCL19, VL19] aim to solve the challenging 3D-MPPE problem where both root-relative poses and absolute root locations need estimating. To that end, many additional constraints [MSM+20, ZMS18] are introduced to infer the absolute depth of individuals, which lead to extra post-processing. Otherwise, the method [RWS17] estimates the root depth by directly minimizing the re-projection errors of the predicted 3D poses.

In recent years, some methods [MCL19, VL19] have made efforts to solve the 3D-MPPE problem by end-to-end learning techniques. The approach in [MCL19] adopts the top-down framework which is proven effective with both 2D- [FXTL17, LWZ+19, CHS+19] and 3D- MPPE [MSM+18, DJH+19, BVGH19] techniques. They present a purely learning-based method and particularly exploit the relationship between root depth and the projection area of one detected person. In contrast, the bottom-up framework is adopted in method [VL19] where a scene depth estimation module and a 2D-MPPE network are employed. In conjunction with these components, the final 3D poses can be effectively estimated by picking the 2D poses and the corresponding depth predictions as inputs.

Referring to the 2D-MPPE task [FXTL17], we apply the top-down structure to solve the 3D-MPPE task. We use the detection network to detect the bounding boxes of humans in an input image. Using the 2D human box detected by the detection network as input, we split the 3D-MPPE task into 3D-SPPE tasks. To solve the task of estimating absolute 3D human pose by a monocular camera, we propose a general framework: the 3D localization of persons, relative 3D human pose estimation.

For the 3D localization of person task, we split the 3D human root coordinate task into root depth estimation and root 2D coordinate estimation. Inspired by the observation in previous research [MCL19] that the root depth of a specific person can be estimated by adjusting her/his projection area with one correction factor, we propose a more effective learning-based approach in this paper. Specifically, we show that the projection area of a detected person can be affected by multiple factors including her/his depth, height, pose, and even the inter-occlusions rather than a single factor. In other words, the formerly proposed correction factor can be decomposed into multiple factors to better estimate one's root depth. Hence, a 3D localization network is designed in this paper to predict these decomposed factors individually. Once these factors are obtained, the depth of a specific person can be calculated on top of the detected bounding boxes due to the fact that it is inversely proportional to the projection area.

For relative 3D human pose estimation task, Our baseline uses [SXW+18] the same way as [MCL19]. This method has achieved state-of-the-art results on some data sets. In some special scenes, the accuracy still needs to be improved, such as complex poses and occlusion scenes. Different from the previous method [SXW+18, MCL19], we propose a multi-scale feature fusion module and introduce an attention mechanism [HZF21] in relative 3D human pose estimation task. This design enables the network to integrate multi-scale information during the up-sampling process while enhancing effective information and suppressing invalid information.

The proposed method is evaluated on several widely adopted datasets, for both single-person (Human3.6m) and multi-person (MuPoTS-3D), against the state-of-the-art baselines. As shown in experiments, our method significantly improves the localization accuracy of human root joints. Moreover, as the root-relative pose is represented by image coordinates and depth, more accurate root depth can also improve the performance of pose estimation. In addition, our framework is more flexible because that the proposed 3D localization network is compatible with most of the existing top-down 2D/3D MPPE/SPPE methods.

3.2 Related Works

Relative 3D Human Pose Estimation. Most 3D pose predictors estimate the body joint coordinates relative to a root joint, usually the pelvis. Relative 3D human pose estimation is divided into two tasks: 3D-SPPE and 3D-MPPE. 3D-SPPE tasks are usually divided into single and two-stage frameworks. Single-stage methods directly estimates 3D human pose through a single RGB image. The two-stage method first obtains the pixel coordinates of the 2D human pose joint points through the 2D human pose estimator and then estimates the 3D human body joint points. 3D-MPPE tasks are usually divided into Top-down and Bottom-up frameworks, which are the same as 2D-MPPE tasks.

For the single-stage 3D-SPPE tasks, [LC14] applied deep learning to the task of 3D human pose estimation for the first time. [PZDD17] applied Stacked Hourglass Networks [NYD16] to 3D human pose to predict a 3D heatmap instead of single coordinates. For the two-stage 3D-SPPE tasks, Some methods [FXW+18, MHRL17, LLL18, YOW+18, ZHS+17b, CL18, CR17] propose to divide the 3D pose estimation task into two parts: first estimate the 2D human pose and then estimate the 3D human pose based on the 2D human pose.

For the Top-Down 3D-MPPE tasks, [RWS17] proposed LCR-Net, which consists of localization, classification, and regression parts. First, the localization part detects the human body bounding box. Then the classification part classifies humans into several anchor poses. Finally, the regression part corrects the difference between the joint points of the regression template and the ground truth. [MCL19] also proposed a top-down scheme which mainly includes two parts: 2D human detector and single-stage 3D-SPPE network. For the Bottom-Up 3D-MPPE tasks, [MSM+18] introduced an occlusion-robust pose-map formulation which supports pose inference for more than one person through PAFs [CHS+19]. Interested readers may refer to the survey paper [ZWC+23, SBIK16] and their references.

3D Human Root Localization. For the 3D localization of person task, researchers usually focus on the root depth estimation of the human body in the camera coordinate system. [VL19] combined the 2D human pose estimator and the monocular depth estimation results to obtain human root localization in the camera coordinate system. Openpose [CHS+19] is used as a 2D pose estimator and Megadepth [LS18] is used as a monocular depth estimator. This method makes human root depth dependent on the accuracy and application scenarios of Megadepth [LS18] and does not take advantage of the characteristics of human body structure. [MCL19] proposed an absolute 3D human pose estimation pipeline consists of human detection, absolute 3D human root localization, and relative 3D single-person pose estimation modules. For the 3D human root localization task, [MCL19] introduced a distance measure based on the principle of camera imaging, and designed a network to fine-tune this parameter.

Absolute 3D Human Pose Estimation. [RWS17, MSM+18] estimated the 3D location of the human root by minimizing the distance between the estimated 2D pose and projected 3D pose. However, this design cannot be generalized to relative 3D pose estimation tasks. As their methods do not output relative 3D human pose, such a distance minimization design cannot be used. The pipeline proposed by [MCL19] consists of human detection, absolute 3D human root localization, and root-relative 3D single-person pose estimation modules. This model makes it to extend the 3D single-person pose estimation techniques to the absolute 3D pose estimation of multiple persons without any ground truth information. [VL19] also proposed a method based on the Bottom-Up framework, which mainly includes three parts: 2D human pose estimator, 3D depth estimator, and 3D pose estimation model. First, use the 2D human pose predictor Openpose [CHS+19] to estimate the 2D coordinates of the human joint points, then use the monocular depth predictor MegaDepth [LS18] to estimate the depth by pixel, and finally estimate the absolute

coordinates of the human root in the camera coordinate system. For the relative 3D human pose estimation task, many methods [FXW+18, MHRL17, LLL18, YOW+18, ZHS+17b, CL18, CR17] have estimated the high-precision 3D human skeleton. This shows that it is necessary to propose a 3D multi-person localization method to extend to relative 3D human pose and to estimate absolute 3D human pose instead of true value information.

3.3 Method

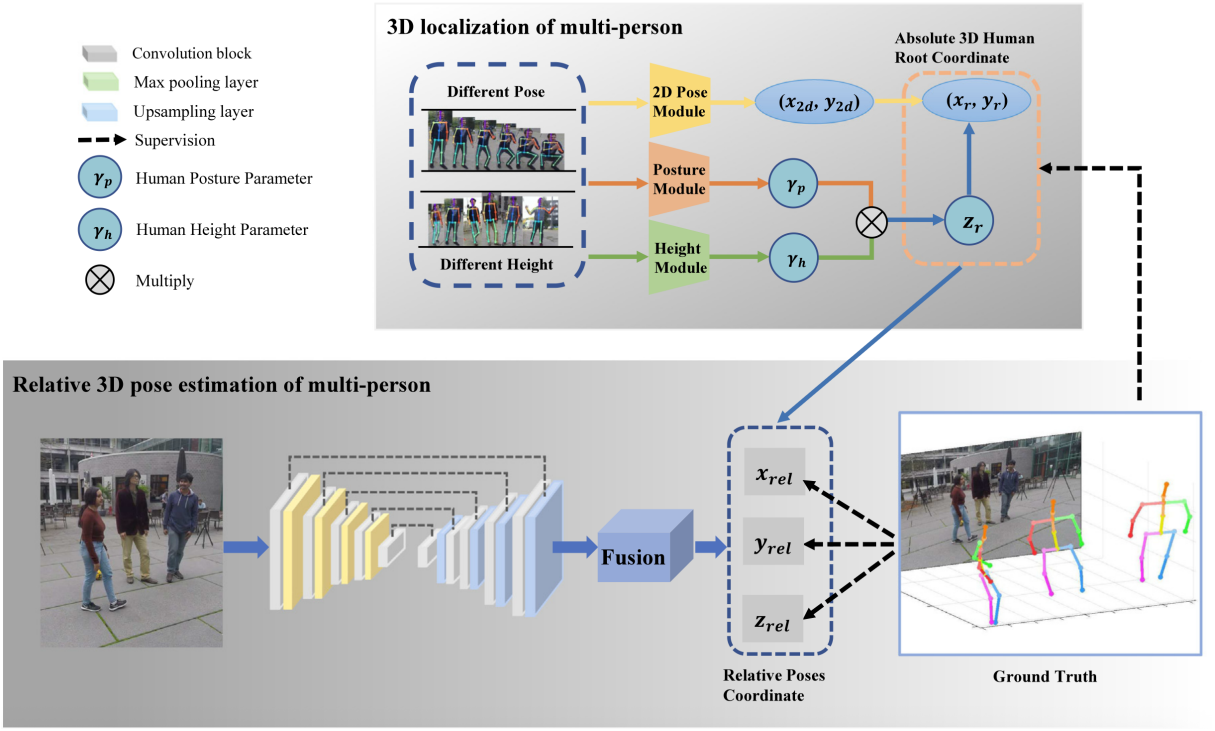


Figure 3.1: **The overall pipeline** consists of two parts: 3D localization of multi-person and relative 3D pose estimation of multi-person. The proposed framework can recover the absolute camera-centred coordinates of multiple persons' key points.

The aim of our method is to estimate the absolute 3D poses of multiple persons from a monocular image. Our approach adopts the top-down framework which consists of the detection, the relative 3D human pose estimation, and the 3D localization networks, illustrated in Fig. 3.1. Within the bounding boxes yielded by the detection network, the pose estimation and the 3D localization networks predict the root-relative poses and absolute coordinates of the human root respectively.

3.3.1 3D Localization Network

In this subsection, we focus mainly on the task of 3D localization. A detailed description of other parts can be found in implementation details section 3.4.3.

The goal of the 3D localization network is to accurately predict the absolute position of human roots in the camera coordinate system. For a person with a static pose, the further he/she stands, the smaller the area he/she projects on the image will be. In other words, one's

Figure 3.2: Different factors that affect the projection area of persons in the image: heights, poses, depths and occlusions.

root depth is inversely proportional to one's projection area if other factors which can affect the projection are fixed. Our approach assumes that one's projection area can be approximated by the area of the corresponding bounding box. An optional strategy [MCL19] is to model all the other factors as a correction factor and then learn a mapping from the input image to the correction. Once the correction factor is predicted by the neural network, the root depth can be obtained conventionally by multiplying the factor to the area of the detected bounding box.

Though being effective to some extent, the performance of the aforementioned strategy is limited due to the fact that there are several (rather than single) factors that can affect the area of projection. As illustrated in Fig. 3.2, besides the depth, there are two other factors that can affect the projection area of a given person: the height and the pose. Therefore, we design a 3D localization network in which the impacts from different sources are disentangled into two branches: a height module and a pose module. In this setting, the learning tasks assigned to each branch become relatively easier: to predict separate correction factors γ_h and γ_p.

Given the predicted correction factors γ_h and γ_p, the root depth z_r can be estimated as follows:

$$z_r = \gamma_h \cdot \gamma_p \cdot \sqrt{f_x f_y (a_{real} / a_{img})}, \tag{3.1}$$

where z_r is the depth of the human root. γ_h and γ_p are the correction factors corresponding to heights and poses respectively. f_x and f_y are the focal lengths divided by per-pixel distance in

x- and y- axes. a_{img} is the projection area approximated by the area of bounding boxes in the image space (pixel^2). a_{real} is a constant representing the area in real space (mm^2), of a person whose height is $h_{template}$ (mm) with a specific pose. In the remainder of this section, we will further present the height and pose modules of our 3D localization network, and the strategy to tackle the occlusions.

3.3.1.1 *Height Module*

The height module estimates the persons' heights from a single image regardless of the poses they are performing. Fig. 3.2 illustrates that in a static pose (e.g. standing), subjects with different heights have different projection areas on the image. Inversely, the heights of the subjects in the real space can, to some extent, be represented by the projection areas on the image, and can thus be estimated from the image. Based on this observation, a height factor is defined to assist the estimation of the depth of each person, as follows:

$$\gamma_h = \frac{h}{h_{ref}}, \tag{3.2}$$

where h is the height of a person in the real space and h_{ref} is a constant which serves as the reference. Given the 3D coordinates of a human joint, we can obtain the ground truth height factor γ_h^*, by summing up the lengths of all corresponding bones, which serves as the supervision.

3.3.1.2 *Pose Module*

Since pose is another factor that affects the projection area, we propose a pose module that focuses merely on modelling the impact of different poses, without considering the variations on the persons' heights. This module estimates a pose factor γ_p which is decoupled from the aforementioned height factor. In this way, the pose module learns the mapping that is invariant to human heights. From Eq. 3.1, the ground truth pose factor γ_p^* can be calculated as:

$$\gamma_p^* = \frac{z_r}{\gamma_h^* \cdot \sqrt{f_x f_y (a_{real} / a_{img})}}. \tag{3.3}$$

Given the ground truth γ_h^* and γ_p^*, both the height module and pose module are trained with L1 loss.

3.3.1.3 *Tackling Occlusions*

When estimating 3D human poses for multiple persons, one inherent problem one needs to tackle is inter-occlusion. Parts of human bodies usually occlude each other in the crowd. As a result, the yielded bounding boxes will be incomplete, which leads to inaccurate correction factors. Rather than designing a module that predicts an additional correction factor, we exploit the pose module to tackle inter-occlusions. Since the pose module maps the poses to corresponding γ_ps, we can add the occluded samples into training data and treat them as some specific pose patterns. Then the yielded γ_p is expected to remove the negative effects caused by the inter-occlusion from the root depth estimation. Since the occlusions usually lead to the shrinkage of the bounding boxes, for randomly selected samples, we add synthetic occlusions to the training samples by randomly moving the boundaries to their centres. The moving distances are controlled by four random variables. The proposed strategy is proven to be effective in our ablation study, illustrating empirically that this method produces more accurate predictions even in the challenging occluded cases.

3.3.2 Root-relative 3D Pose Estimation Network

In this subsection, we focus mainly on the task of root-relative 3D multi-person pose estimation. We follow a two-stage top-down framework similar to in [MCL19]. For human detection, we use Mask-RCNN [HGDG17] as a detection model. We use ResNet [HZRS16] as the backbone network. Based on the attention mechanism, we designed an information fusion module with three consecutive deconvolutional layers and an attention module. Different from the previous work [SXW+18, MCL19], our solution introduces the fusion module to capture the position information and the relationship information between the channels. The pose estimation part takes a feature map from the backbone part and fusion module, then we use a 1-by-1 convolutional layer to generate 3D heatmaps for each joint. We use the soft-argmax operation to extract the 2D image coordinates and the root-relative depth values. We train the root-relative 3D pose estimation network by minimising the L1 distance between the estimated and ground truth coordinates in format (u, v, z) (image coordinates and relative depth).

3.4 Experiments

Table 3.1: MPJPE_{abs} and MPJPE_{rel} comparisons with [MCL19] on Human3.6m (protocol 2), using the groundtruth (GT) and the detected (X-101-32) bounding boxes. Lower is better.

	Absolute joint positions (MPJPE_{abs})				Relative joint positions (MPJPE_{rel})			
Methods	[MCL19]	**ours**	[MCL19]	**ours**	[MCL19]	**ours**	[MCL19]	**ours**
DetNet	X-101-32	X-101-32	GT	GT	X-101-32	X-101-32	GT	GT
Dir.	86.2	**67.6**	108.3	**68.2**	51.3	**51.1**	51.5	**50.7**
Dis.	106.2	**77.2**	113.0	**74.5**	57.0	**56.5**	56.8	**55.9**
Eat	133.0	**103.4**	121.0	**96.0**	51.5	**51.1**	51.2	**50.9**
Gre.	110.2	**92.3**	116.8	**92.4**	52.1	**52.0**	52.2	**51.9**
Phon.	133.0	**114.7**	115.8	**100.6**	56.7	**56.4**	55.2	**54.9**
Pose	88.5	**68.6**	102.7	**68.1**	**47.7**	**47.7**	47.7	**47.2**
Pur.	109.7	**84.2**	114.4	**90.0**	50.8	**50.6**	50.9	**50.4**
Sit	176.3	**135.9**	139.9	**119.3**	64.7	**64.0**	63.3	**62.9**
SitD.	263.5	**173.1**	257.9	**163.8**	69.2	**67.7**	69.9	**68.4**
Smo.	134.0	**108.5**	126.6	**91.8**	55.2	**54.7**	54.2	**53.5**
Phot.	117.0	**80.8**	123.8	**82.0**	57.1	**56.3**	57.4	**56.4**
Wait	103.5	**94.1**	109.7	**90.6**	50.6	**50.5**	50.4	**50.1**
Walk	83.6	**62.6**	99.9	**63.9**	**42.4**	**42.4**	42.5	**42.1**
WalkD.	115.3	**89.9**	127.7	**91.2**	56.7	**56.2**	57.5	**56.9**
WalkP.	86.2	**66.1**	100.5	**67.0**	**47.6**	**47.6**	47.7	**47.3**
Avg.	125.3	**97.0**	125.3	**91.4**	54.7	**54.3**	54.4	**53.8**

3.4.1 Datasets

Human3.6m Dataset To test our method's effectiveness in predicting the absolute root depth from single-person images, we employ Human3.6m [IPOS13], which is a publicly available single-person 3D human pose estimation dataset containing 3.6 million frames of video sequences captured in an indoor environment. The sequences contain 11 professional actors performing 15

Table 3.2: MPJPE_{rel} without using root ground truth on Human3.6m dataset. Lower is better.

Methods	MPJPE_{rel} (Protocol 1)	MPJPE_{rel} (Protocol 2)
[RWS17]	-	87.7
[MHRL17]	45.5	62.9
[YOW+18]	-	58.9
[ZPT+19]	43.8	57.6
[SVB+19]	-	58.0
[MRC+17]	-	69.9
[RWS19]	42.7	63.5
[MCL19]	35.2	54.4
[LKP+20]	34.5	**50.9**
ours	**33.8**	52.7

Table 3.3: MPJPE_{rel} comparisons on Human3.6m under *protocol 1*, using the absolute 3D human root localisation with ground truth. Lower is better.

	[YIK+16]	[CR17]	[MN17]	[ZZP+18]	[MHRL17]	[SXW+18]	[MCL19]	**ours**
Dir.	88.4	71.6	67.4	47.9	39.5	42.1	31.0	**30.2**
Dis.	72.5	66.6	63.8	48.8	43.2	44.3	**30.6**	30.7
Eat	108.5	74.4	87.2	52.7	46.4	45.0	39.9	**39.8**
Gre.	110.2	79.1	73.9	55.0	47.0	45.4	35.5	**35.1**
Phon.	97.1	70.1	71.5	56.8	51.0	51.5	34.8	**34.3**
Pose	81.6	67.6	69.9	49.0	41.4	43.2	**30.2**	31.2
Pur.	107.2	89.3	65.1	45.5	40.6	41.3	32.1	**31.0**
Sit	119.0	90.7	71.7	60.8	56.5	59.3	35.0	**34.3**
SitD.	170.8	195.6	98.6	81.1	69.4	73.3	43.8	**41.5**
Smo.	142.5	83.5	81.3	53.7	49.2	51.0	35.7	**35.2**
Phot.	145.2	93.3	93.3	65.5	56.0	53.0	37.6	**37.3**
Wait	86.9	72.1	74.6	51.6	45.0	44.0	30.1	**29.6**
Walk	92.1	55.7	76.5	50.4	38.0	38.3	24.6	**24.2**
WalkD.	165.7	85.9	77.7	54.8	49.5	48.0	**35.7**	35.8
WalkP.	165.7	62.5	74.6	55.9	43.1	44.8	29.3	**28.0**
Avg.	108.3	82.7	76.5	55.3	47.7	48.3	34.0	**33.5**

activities. The training set of the dataset includes 7 subjects (S1, 5, 6, 7, 8, 9, and 11), all of which are annotated with ground truth 3D poses. We adopt standard protocols employed in most of the previous works [MCL19, VL19], using a subset of the training set to train and the rest of the training set for testing, as well as downsampling the training subset of the videos by every 5 frames and the testing subset by every 64 frames. **Protocol 1** uses six subjects (S1, S5, S6, S7, S8, S9) in training and S11 in testing. **Protocol 2** uses five subjects (S1, S5, S6, S7, S8) in training and two subjects (S9, S11) in testing. If not specifically stated, **protocol 2** is usually used as a generic protocol for Human3.6m. Specifically, no additional 2D human pose dataset is exploited to assist the training of our network.

MuCo-3DHP and MuPoTS-3D Datasets For the multi-person case, we exploit two 3D multi-person pose estimation datasets, namely MuCo-3DHP and MuPoTS-3D, proposed in [MSM+18], for training and testing respectively. MuCo-3DHP is a synthetic dataset, generated by composing the humans captured in the existing MPI-INF-3DHP 3D single-person pose estimation dataset [MRC+17]. The MuPoTS-3D dataset contains 8000 frames captured in both indoor and outdoor scenes, with challenging occlusions and person-person interactions. The ground truth 3D poses

Table 3.4: MPJPE_{rel} comparisons on Human3.6m under *protocol 2*, using the absolute 3D human root localisation with ground truth. Lower is better.

	[CR17]	[TRA17]	[MN17]	[ZZP+18]	[JY17]	[MRC+17]	[MHRL17]	[FXW+18]	[SSLW17]	[SXW+18]	[MCL19]	**ours**
Dir.	89.9	65.0	69.5	68.7	74.4	57.5	51.8	50.1	52.8	**47.5**	50.5	48.6
Dis.	97.6	73.5	80.2	74.8	66.7	68.6	56.2	54.3	54.8	**47.7**	55.7	52.6
Eat	90.0	76.8	78.2	67.8	67.9	59.6	58.1	57.0	54.2	49.5	50.1	**49.3**
Gre.	107.9	86.4	87.0	76.4	75.2	67.3	59.0	57.1	54.3	**50.2**	51.7	51.0
Phon.	107.3	86.3	100.8	76.3	77.3	78.1	69.5	66.6	61.8	**51.4**	53.9	52.7
Pose	93.6	68.9	76.0	84.0	70.6	56.9	55.2	53.4	53.1	**43.8**	46.8	46.1
Pur.	136.1	74.8	69.7	70.2	64.5	69.1	58.1	55.7	53.6	**46.4**	50.0	49.9
Sit	133.1	110.2	104.7	88.0	95.6	98.0	74.0	72.8	71.7	**58.9**	61.9	62.9
SitD.	240.1	173.9	113.9	113.8	127.3	117.5	94.6	88.6	86.7	65.7	68.0	**63.9**
Smo.	106.7	85.0	89.7	78.0	79.6	69.5	62.3	60.3	61.5	**49.4**	52.5	52.0
Phot.	139.2	110.7	110.7	98.4	79.1	82.4	78.4	73.3	67.2	55.8	55.9	**54.7**
Wait	106.2	85.8	85.8	90.1	73.4	68.0	59.1	57.7	53.4	**47.8**	49.9	48.2
Walk	87.0	71.4	71.4	62.6	67.4	55.3	49.5	47.5	47.1	**38.9**	41.8	41.5
WalkD.	114.1	86.3	86.3	75.1	71.8	76.5	65.1	62.7	61.6	**49.0**	56.1	53.3
WalkP.	90.6	73.1	73.1	73.6	72.8	61.4	52.4	50.6	63.4	**43.8**	46.9	46.7
Avg.	114.2	88.4	88.4	79.9	77.6	72.9	62.9	60.4	59.1	**49.6**	53.3	52.0

Table 3.5: MRPE comparisons with [VL19] and [MCL19] on Human3.6m and MuPoTS-3D datasets, using the ground truth (GT) and the detected (X-101-32) bounding boxes. As the root depth estimation is affected by the bounding box detection, more accurate prediction can be obtained when using ground-truth bounding boxes.

Methods	DetecNet	Human3.6m	MuPoTS-3D
[VL19]	-	-	274
[MCL19]	X-101-32	120	286
[MCL19]	GT	121	272
ours	X-101-32	**95.7**	**253**
ours	GT	**88.5**	**225**

of this dataset are obtained with a multi-view marker-less motion capture system. Following previous protocol [MCL19, VL19], we perform background augmentation to half of the frames in the MuCo-3DHP dataset. The background augmentation is done utilising the images (without humans) from the COCO dataset [LMB+14]. Again, when training our 3D localisation network on the MuCo-3DHP dataset, we do not employ any additional 2D human pose dataset.

3.4.2 Evaluation Metrics

We employ the commonly adopted mean per joint position error (MPJPE) as a metric to evaluate the accuracy of the predicted 3D poses (the predicted 3D poses here stand for the joint coordinates in camera system which are converted from (u, v, z) with a given root depth). The original MPJPE, however, is calculated after aligning the 3D poses to the position of the human root joint (pelvis). The root joint alignment operation makes the traditional MPJPE metric unable to evaluate the accuracy of the absolute 3D positions of joints. Toward this end, to evaluate the

Table 3.6: Sequence-wise 3DPCK$_{rel}$ comparisons with [VL19] and [MCL19] on the MuPoTS-3D dataset, using the ground truth (GT) and the detected (X-101-32) bounding boxes. Higher is better.

	Accuracy for all ground truths.				Accuracy for only matched ground truths.				
Methods	[MCL19]	**ours**	[MCL19]	**ours**	[VL19]	[MCL19]	**ours**	[MCL19]	**ours**
DetNet	X-101-32	X-101-32	GT	GT	-	X-101-32	X-101-32	GT	GT
S1	59.5	**68.2**	51.7	**72.7**	55.4	59.5	**68.2**	52.1	**73.2**
S2	**44.7**	40.3	43.8	**44.7**	21.9	**45.3**	41.4	48.2	**49.1**
S3	51.4	**53.3**	**53.9**	42.5	38.3	51.4	**53.4**	**53.9**	42.5
S4	46.0	**58.5**	48.3	**57.9**	8.2	46.2	**59.0**	48.3	**57.9**
S5	52.2	**59.2**	51.2	**64.1**	52.8	53.0	**60.0**	51.7	**64.8**
S6	**27.4**	26.2	33.3	**38.1**	9.7	**27.4**	26.2	33.3	**38.2**
S7	**23.7**	14.1	**33.5**	27.7	14.5	**23.7**	14.2	**33.5**	27.7
S8	26.4	**51.3**	34.7	**57.8**	13.7	26.4	**51.3**	38.6	**64.4**
S9	39.1	**56.5**	50.9	**65.7**	27.7	39.1	**56.5**	50.9	**65.7**
S10	23.6	**67.4**	24.5	**81.4**	60.4	23.6	**67.4**	24.5	**81.4**
S11	18.3	**31.0**	19.9	**30.1**	22.0	18.3	**31.0**	19.9	**30.1**
S12	14.9	**29.0**	7.3	**19.0**	27.1	14.9	**29.0**	7.3	**19.0**
S13	38.2	**42.1**	37.7	**44.3**	18.5	38.2	**42.4**	38.2	**44.8**
S14	26.5	**26.8**	18.3	**34.1**	24.3	29.5	**30.3**	18.9	**35.2**
S15	**36.8**	32.1	30.3	**31.5**	**39.1**	36.8	32.1	30.3	**31.5**
S16	23.4	**29.3**	20.7	**26.8**	**29.7**	23.6	29.6	20.7	**26.8**
S17	14.4	**20.8**	9.5	**28.5**	**35.9**	14.4	20.8	9.5	**28.5**
S18	19.7	**30.5**	19.6	**31.2**	27.7	20.0	**30.9**	19.6	**31.2**
S19	18.8	**22.1**	16.9	**20.3**	**32.0**	18.8	22.1	17.1	**20.6**
S20	25.1	**26.0**	22.0	**27.4**	**43.8**	25.4	26.7	23.8	**29.7**
Avg.	31.5	**39.2**	31.4	**39.6**	30.1	31.8	**42.3**	32.0	**43.1**

performance of estimating the absolute joint positions, we define **MPJPE**$_{abs}$, representing the absolute MPJPE without performing root alignment. To distinguish between these two metrics, we rename the original MPJPE to **MPJPE**$_{rel}$, declaring that it is relative to the root joint.

We also adopt the 3D percentage of correct key points (3DPCK) proposed in [MRC$^+$17] as an evaluation metric for 3D poses. In the setting of our experiments, a predicted joint position is regarded as correct if its distance toward the ground truth falls within a threshold of 15 cm. Again, we use the notation **3DPCK**$_{abs}$ for the absolute 3DPCK without root alignment to evaluate predictions represented in absolute camera-centred coordinates, and rename the original 3DPCK to **3DPCK**$_{rel}$.

To evaluate the accuracy of the estimated 3D human roots, we utilise mean of the root position error (**MRPE**) introduced by [MCL19]. It is defined as the mean of Euclidean distances between the coordinates of the estimated root joints and the ground truth. Another metric for evaluating the absolute root position is the average precision of 3D human root location (**AP**$^{root}_{25}$). This metric assumes a prediction to be correct if the Euclidean distance between the estimated and the ground truth coordinates is less than 25cm.

3.4.3 Implementation Details

Our model contains two sub-network, localisation network and pose estimation network. For the task of 3D person localisation, the input images are resized to 256×256 pixels and fed into the network. Firstly, we use DetectNet to obtain bounding boxes from the input images. The

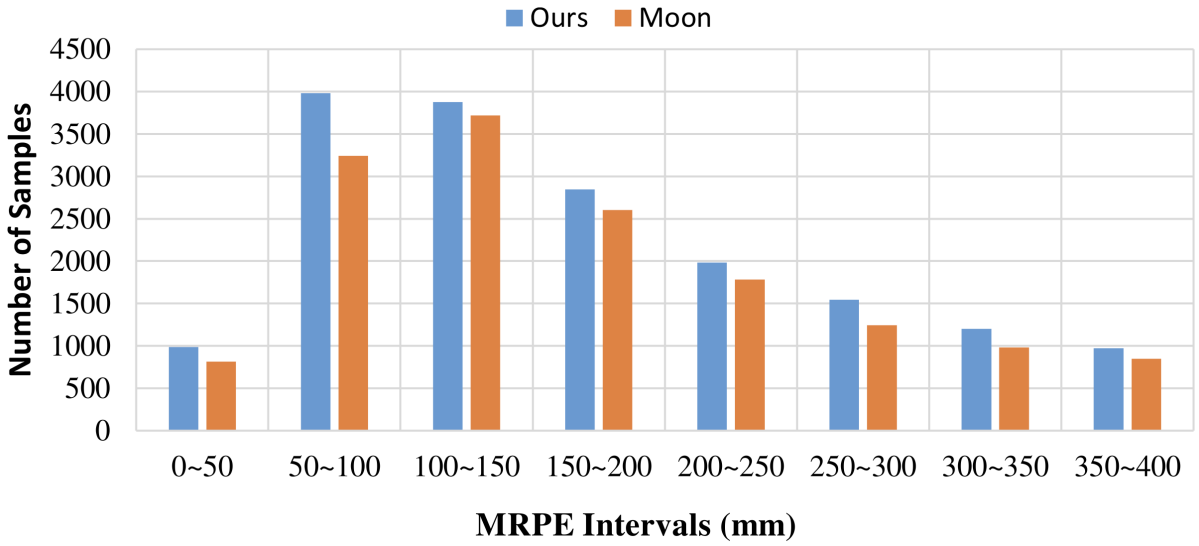


Figure 3.3: Number of estimations that fall into different MRPE intervals on MuPoTS-3D dataset. The more number of the samples fall into low MRPE intervals, the better the performance. We compare our method (ours) with the state-of-the-art method (Moon [MCL19]).

DetectNet is constructed by a Mask R-CNN [HGDG17] with X-101-32 backbone [XGD+17] and pre-trained on the COCO dataset. Then, the bounding boxes can help cropping the input images. We fed the cropped patches into our 3D localization network, which composed of a backbone and a depth network simultaneously. The backbone extracts $2048 \times 8 \times 8$ global features of the input image with a ResNet [HZRS16]. The depth network consists of the pose module and the height module. These two modules extracts the image features by using the Resnet [HZRS16] backbone and then perform global average pooling to obtain feature map of $2048 \times 1 \times 1$. According to the down-sampled feature map, the pose factor γ_p and the height factor γ_h are computed by a 1×1 convolution layer. For relative 3D human pose estimation task, we proposed a root-relative 3D pose estimation network, incorporating the CA attention mechanism [HZF21], to predict root-relative coordinates of body joints.

Our 3D localization network is implemented with PyTorch. We initialize the backbone module with the publicly released ResNet-50 [HZRS16] pre-trained on the ImageNet dataset [RDS+15]. Adam optimiser [KB15] is used to update the weights with a min-batch of size 128. We set the initial learning rate to 1×10^{-1}. The cropped images are resized to 256×256 before being input into the backbone module. During training, we exploit data augmentation strategies including rotation ($\pm 30°$) and horizontal flipping.

When training the pose module, we randomly select 20% of the bounding box cropped input images to be shrunk to simulate the case when occlusions occur. In specific, based on the observation that the lower part of the human body is more likely to get occluded, we shrink each selected image by three random values between $[0, 10\%]$ for the top, left, and right margins respectively, and another random value between $[0, 50\%]$ for the bottom margin. The height module is trained independently from the pose module. We train the pose module for 20 epochs on two NVIDIA 1080ti GPUs in 25 h. For the height module, 12 h are spent for 10 epochs on two NVIDIA 1080ti GPUs.

Figure 3.4: **Qualitative results** of the absolute 3D human pose on MuPoTS-3D, shown in 3D space.

3.4.4 Comparison with State-of-the-art Methods

We test our method on both Human3.6m and MuPoTS-3D datasets and compare our results with the results produced by state-of-the-art methods [MCL19, VL19]. Both methods are end-to-end, achieving state-of-the-art results on estimating the absolute 3D human poses. [MCL19] adopt a top-down framework to estimate human depth and 3D absolute human joint in camera coordinates and achieve the (former) best results on the Human3.6m dataset. [VL19] employ a bottom-up framework, exploiting a scene depth estimation module and a 3D-MPPE network, and obtains state-of-the-art results on the MuPoTS-3D dataset. We follow their experimental configurations to test our network with and without ground truth bounding boxes for fair comparisons.

3.4.4.1 *Results on Human3.6m*

For estimating the absolute localization of the human root joints in the single-person case, our method achieves significantly better results than the former state-of-the-art method [MCL19]

Table 3.7: Sequence-wise 3DPCK$_{rel}$ comparisons with the previous methods on the MuPoTS-3D dataset, using the groundtruth (GT) and the detected (X-101-32) bounding boxes. Higher is better.

	Accuracy for all ground truths.					Accuracy for only matched ground truths.				
	[RWS17]	[MRC$^+$17]	[RWS19]	[MCL19]	**ours**	[RWS17]	[MRC$^+$17]	[RWS19]	[MCL19]	**ours**
S1	67.7	81.0	87.3	94.4	**95.6**	69.1	81.0	88.0	94.4	**96.3**
S2	49.8	60.9	61.9	77.5	**78.2**	67.3	65.3	73.3	78.6	**80.1**
S3	53.4	64.4	67.9	**79.0**	78.5	54.6	64.6	67.9	79.0	**79.8**
S4	59.1	63.0	74.6	**81.9**	80.0	61.7	63.9	74.6	**82.1**	81.2
S5	67.5	69.1	78.8	85.3	**86.0**	74.5	75.0	81.8	86.6	**86.8**
S6	22.8	30.3	48.9	72.8	**75.0**	25.2	30.3	50.1	72.8	**75.4**
S7	43.7	65.0	58.3	81.9	**82.3**	48.4	65.1	60.6	81.9	**82.7**
S8	49.9	59.6	59.7	75.7	**78.1**	63.3	61.1	60.8	75.8	**78.4**
S9	31.1	64.1	78.1	90.2	**92.9**	69.0	64.1	78.2	90.2	**93.1**
S10	78.1	83.9	89.5	**90.4**	**90.4**	78.1	83.9	89.5	90.4	**90.9**
S11	50.2	68.0	69.2	79.2	**80.2**	53.8	72.4	70.8	79.4	**80.7**
S12	51.0	68.6	73.8	79.9	**80.6**	52.2	69.9	74.4	79.9	**81.8**
S13	51.6	62.3	66.2	75.1	**76.6**	60.5	71.0	72.8	75.3	**78.2**
S14	49.3	59.2	56.0	**72.7**	71.0	60.9	72.9	64.5	**81.0**	72.9
S15	56.2	70.1	74.1	81.1	**82.9**	59.1	71.3	74.2	81.0	**83.5**
S16	66.5	80.0	82.1	89.9	**90.0**	70.5	83.6	84.9	90.7	**91.5**
S17	65.2	79.6	78.1	**89.6**	89.3	76.0	79.6	85.2	89.6	**90.2**
S18	62.9	67.3	72.6	**81.8**	81.0	70.0	73.5	78.4	**83.1**	82.8
S19	66.1	66.6	73.1	81.7	**82.9**	77.1	78.9	75.8	81.7	**83.4**
S20	59.1	67.2	61.0	76.2	**78.3**	81.4	90.9	74.4	77.3	**79.5**
Avg.	53.8	66.0	70.6	81.8	**82.5**	62.4	70.8	74.0	82.5	**83.5**

on the Human3.6m dataset. We compare our results generated with and without ground truth bounding boxes. When using the detected bounding boxes, we adopt the same DetectNet as is used in [MCL19], namely the Mask R-CNN [HGDG17] with X-101-32 backbone [XGD$^+$17] pre-trained on the COCO dataset, for fair comparisons. Note that unlike [MCL19], we do not exploit any additional human pose dataset to assist the training of our network, which means that our method is effective even in more challenging training scenarios.

In terms of the absolute depth estimation, i.e. the MRPE metric, our method gains a considerable 20.5% improvement over the results of [MCL19] (the 2nd and the 4th row in Table 3.5) with detected bounding boxes and 26.3% with ground truth bounding boxes (the 3rd and the 5th row in Table 3.5). The results and comparisons for the MPJPE metric are demonstrated in Table 3.1. It is shown that our method improves the estimated absolute positions of 3D poses (MPJPE$_{abs}$) with and without ground truth bounding boxes by 22.6% and 27.0%, respectively. In addition, it can also be observed from Table 3.1 that in terms of root-relative 3D poses (MPJPE$_{rel}$), our method still gains slight improvements over that of [MCL19] with and without ground truth bounding boxes, thanks to our more accurate estimations of the absolute root joints. In fact, with the very PoseNet adopted in [MCL19] and our method, MPJPE$_{rel}$ saturates at 53.3mm (average of all the actions) on H36M using the ground truth root joint positions to perform root alignment. Our results (53.8mm) is already very close to this lower limit.

For root-relative 3D human pose estimation task using root ground truth in MPJPE$_{rel}$ matrix, our method is better than the previous state-of-the-art method [MCL19] in protocol 1 (Table 3.3), and is better than our baseline method in protocol 2 (Table 3.4). Without using ground truth in Table 3.2, our method is better than the previous state-of-the-art method [LKP$^+$20] in protocol 1, and is better than our baseline method in protocol 2.

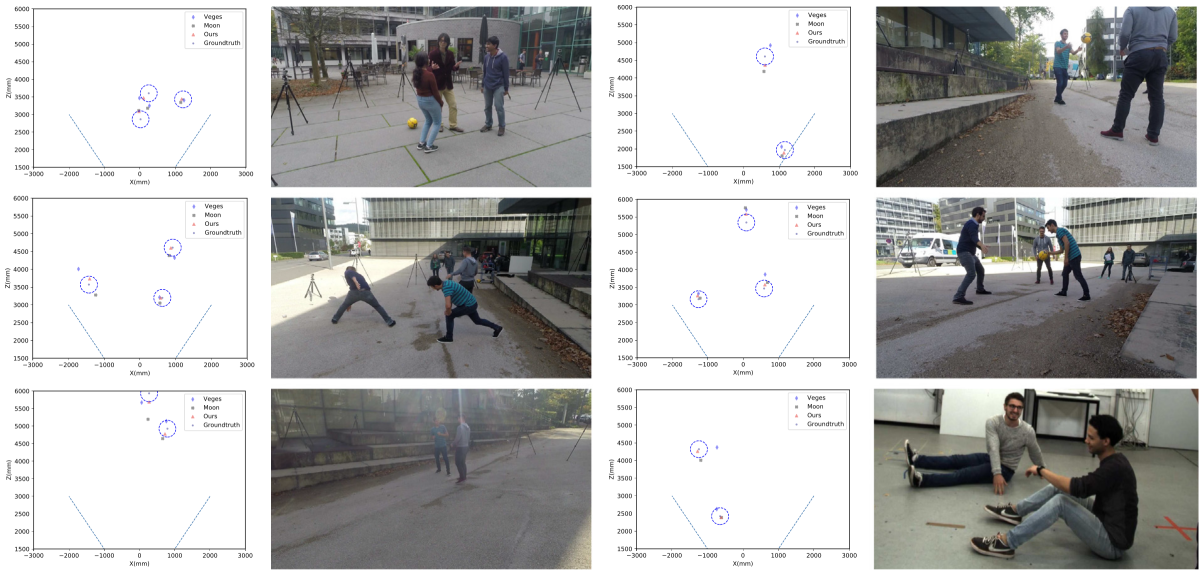


Figure 3.5: **Qualitative results** and comparisons of the absolute root localisation on MuPoTS-3D, shown in bird's eye view.

Table 3.8: Ablation study. We studied the influence of occ (tackling occlusions) and CA (CA attention) factors on our method with metrics AP_{25}^{root} on MuPoTS-3D (higher is better) and $MPJPE_{rel}$ on Human3.6m (lower is better), where '-' means removing specific components.

	AP_{25}^{root}	$MPJPE_{rel}$
	gt detection / person detection	root gt / root estimation
baseline	31.4 / 31.0	53.3 / 54.4
ours-occ-CA	37.9 / 36.5	54.0 / 54.1
ours-CA	**44.1 / 39.8**	53.3 / 53.8
ours	**44.1 / 39.8**	**52.0 / 52.7**

3.4.4.2 *Results on MuPoTS-3D*

In the case of 3D multi-person root estimation, we also achieve state-of-the-art results. The last column of Table 3.5 shows that our method obtains significantly better MRPE on the MuPoTS-3D dataset than both of former state-of-the-art, multi-person root estimation methods [VL19] and [MCL19]. As illustrated in Fig. 3.3, we obtain more predictions in the lower-MRPE intervals comparing with the baseline methods. In terms of $3DPCK_{abs}$, it is demonstrated in Table 3.6 that our method beats the former state-of-the-arts with or without ground truth bounding boxes, with both of the evaluation protocols provided by the MuPoTS-3D dataset of whether to apply identity matching to the predictions with ground truth visibility. Some qualitative results and comparisons of are displayed in Fig. 3.5, in which the absolute root depth estimations generated by our method are clearly closer to the ground truth. For Root-relative 3D human pose estimation task using root ground truth in $3DPCK_{rel}$, whether to use root ground truth or not, our solution is better than the previous methods in Table 3.7. Some qualitative results of the absolute 3D human pose on MuPoTS-3D are shown in Fig. 3.4.

3.4.5 Ablation Study

We conduct some comparative experiments on the multi-person 3D human pose dataset MUPoTS-3D to verify the impact of various components of our 3D localisation network on the depth estimation, shown in Table 3.8. With the decomposition of the Pose module and the Height module, our method (indicated as ours w/o occ.) gains improvements over the baseline model [MCL19] in both AP_{25}^{root}, with and without using the ground truth bounding boxes, which means that the decomposition really counts. In addition, as we find that generic human root depth prediction methods usually fail on 3D multi-person datasets due to the impact of inter-occlusions, we list our results acquired by both the methods with and without the occlusion tackling operation. The last 4 rows in Table 3.8 show that our method with occlusion tackling achieves further improvements over the one without, demonstrating the effectiveness of our occlusion tackling strategy.

3.5 Conclusion

In this paper, we propose a new model towards the problem of 3D multi-person human pose estimation from a single RGB image, focusing mainly on estimating the absolute depth of humans in multi-person scenes. Based on the assumption that the root depth is inversely proportional to the projection area of a detected person, we acquire the final root depth indirectly by predicting the correction factors. Unlike the previous methods, we decompose the correction factors into two parts: the pose and the height, due to the observation that the area of a detected person is mainly affected by these two factors. To this end, we design a 3D localization network consisting of a pose module and a height module. In addition, to deal with the common problem of inter-occlusion that occurred in multi-person pose estimation datasets, we propose an occlusion tackling strategy, augmenting the bounding boxes used to crop the input images to assist the training of the pose module of our network. For the 3D human relative pose estimation task, we use the attention mechanism to capture the correlation between human joint points and further improve the accuracy of relative pose estimation. The proposed networks achieve state-of-the-art results on human root estimation, performing significantly better than the previous methods. Our method has various important applications, such as localizing the human positions for the task of pedestrian recognition and reconstructing 3D multi-person interaction models.

4 Human Motion Capture from Scene Contact Guidance

Contents

Marker-less monocular 3D human motion capture (MoCap) with scene interactions is a challenging research topic relevant for extended reality, robotics and virtual avatar generation. Due to the inherent depth ambiguity of monocular settings, 3D motions captured with existing methods often contain severe artefacts such as incorrect body-scene inter-penetrations, jitter and body floating. To tackle these issues, we propose HULC, a new approach for 3D human MoCap which is aware of the scene geometry. HULC estimates 3D poses and dense body-environment surface contacts for improved 3D localisations, as well as the absolute scale of the subject. Furthermore, we introduce a 3D pose trajectory optimisation based on a novel pose manifold sampling that resolves erroneous body-environment inter-penetrations. Although the proposed method requires less structured inputs compared to existing scene-aware monocular MoCap algorithms, it produces more physically-plausible poses: HULC significantly and consistently outperforms the existing approaches in various experiments and on different metrics. Project page: `https://vcai.mpi-inf.mpg.de/projects/HULC/`.

This chapter is based on [SGL+22]. As the third author, Zhi Li contributed to the development and refinement of the dense contact generation components used in this work. Conducted in parallel with Zhi Li's first-author project on human motion capture in deformable environments ([LSS+22], Chapter 5), this paper explores a distinct main idea centred on dense contact guidance. Zhi Li participated in extensive technical discussions that helped stabilise the shared contact-related modules across both projects. In addition, Zhi Li implemented key parts of the GTA-IM data annotation pipeline in this paper (Section 4.4), i.e. the tracking procedure and SMPL parameter optimisation, which form the basis of the training and supervision data used in this study.

4.1 Introduction

3D human motion capture (MoCap) from a single colour camera received a lot of attention over the past years [MSS+17, MSM+20, KBJM18, KZFM19, KAB20, PFGA19, CML21, TKS+16, MRC+17, RSF18, CR17, MHRL17, TRA17, MN17, PZZD18, NYD16, PZDD17, YOW+18, ZHS+17a, HXM+19, BKL+16, WC10, KPBD19, SAA+20, SYL+19, KHHB21, KHT+21, KPJD21].

Table 4.1: Overview of inputs and outputs of different methods. "$\boldsymbol{\tau}$" and "env. contacts" denote global translation and environment contacts, respectively. "$*$" stands for sparse marker contact labels.

Approach	Inputs	Outputs				
		body pose	$\boldsymbol{\tau}$	absolute scale	body contacts	env. contacts
PROX [HCTB19]	RGB + scene mesh	✓	✓	✗	✗	✗
PROX-D [HCTB19]	RGBD + scene mesh	✓	✓	✗	✗	✗
LEMO [ZZB+21]	RGB(D) + scene mesh	✓	✓	✗	✓*	✗
HULC (ours)	RGB + scene point cloud	✓	✓	✓	✓	✓

Its applications range from mixed and augmented reality, to movie production and game development, to immersive virtual communication and telepresence. MoCap techniques that not only focus on humans *in a vacuum* but also account for the scene environment—this encompasses awareness of the physics or constraints due to the underlying scene geometry— are coming increasingly into focus [SGX+21, SGXT20, RGH+20, HCTB19, ZZB+21, ZMS18, RBH+21, YZH+22].

Taking into account interactions between the human and the environment in MoCap poses many challenges, as not only articulations and global translation of the subject must be accurate, but also contacts between the human and the scene need to be plausible. A misestimation of only a few parameters, such as a 3D translation, can lead to reconstruction artefacts that contradict physical reality (*e.g.,* body-environment penetrations or body floating).

On the other hand, known human-scene contacts can serve as reliable boundary conditions for improved 3D pose estimation and localisation. While several algorithms merely consider human interactions with a ground plane [SGX+21, SGXT20, RGH+20, RBH+21, ZMS18], a few other methods also account for the contacts and interactions with the more general 3D environment [HCTB19, ZZB+21]. However, due to the depth ambiguity of the monocular setting, their estimated subject's root translations can be inaccurate, which can create implausible body-environment collisions. Next, they employ a body-environment collision penalty as a soft constraint. Therefore, the convergence of the optimisation to a bad local minima can also cause unnatural body-environment collisions.

This paper addresses the limitations of the current works and proposes a new 3D **HU**man MoCap framework with pose manifold samp**L**ing and guidance by body-scene **C**ontacts, abbreviated as HULC. It improves over other monocular 3D human MoCap methods that consider constraints from 3D scene priors [HCTB19, ZZB+21]. Unlike existing works, HULC estimates contacts not only on the human body surface but also on the environment surface for the improved global 3D translation estimations. Next, HULC introduces a pose manifold sampling-based optimisation to obtain plausible 3D poses while handling the severe body-environment collisions in a *hard manner*. Our approach regresses more accurate 3D motions respecting scene constraints while requiring less-structured inputs (*i.e.,* an RGB image sequence and a point cloud of the static background scene) compared to the related monocular scene-aware methods [HCTB19, ZZB+21] that require a complete mesh and images. HULC returns physically-plausible motions, an absolute scale of the subject and dense contact labels both on a human template surface model and the environment.

HULC features several innovations which in interplay enable its functionality, *i.e.,* 1) a new learned implicit function-based dense contact label estimator for humans and the general 3D scene environment, and 2) a new pose optimiser for scene-aware pose estimation based on a pose manifold sampling policy. The first component allows us to jointly estimate the absolute

subject's scale and its highly accurate root 3D translations. The second component prevents severe body-scene collisions and acts as a hard constraint, in contrast to widely-used soft collision losses [HCTB19, MGT+19]. To train the dense contact estimation networks, we also annotate contact labels on a large scale synthetic daily motion dataset: GTA-IM [CGM+20].

To summarise, our primary technical contributions are as follows:

- A new 3D MoCap framework with simultaneous 3D human pose localisation and body scale estimation guided by estimated contacts. It is the first method that regresses the dense body and environment contact labels from an RGB sequence and a point cloud of the scene using an implicit function (Sec. 4.3.1).
- A new pose optimisation approach with a novel pose manifold sampling yielding better results by imposing hard constraints on incorrect body-environment interactions (Sec. 4.3.2).
- Large-scale body contact annotations on the GTA-IM dataset [CGM+20] that provides synthetic 3D human motions in a variety of scenes (Fig. 4.1 and Sec. 4.4).

We report quantitative results, including an ablative study, which show that HULC outperforms existing methods in 3D accuracy and on physical plausibility metrics (Sec. 4.5).

4.2 Related Works

Most monocular MoCap approaches estimate 3D poses alone or along with the body shape from an input image or video [HXM+19, KBJM18, KZFM19, KAB20][CML21, TKS+16, MRC+17, RSF18][CR17, JKP+20, MHRL17, TRA17][MN17, PZZD18, NYD16, ZHS+17a] [PZDD17, BKL+16, WC10, KHHB21][KPBD19, SAA+20, SYL+19, YOW+18][KPJD21, ZHW20]. Some methods also estimate 3D translation of the subject in addition to the 3D poses [MSS+17, MSM+20, KHT+21, PFGA19]. Fieraru *et al.* [FZO+20] propose a multi-person 3D reconstruction method considering human-human interactions. Another algorithm class incorporates an explicit physics model into MoCap and avoids environmental collisions [SGXT20, RGH+20, SGX+21, YWS+21]. These methods consider interactions with only a flat ground plane or a stick-like object [LSC+19], unlike our HULC, that can work with arbitrary scene geometry.

Awareness of human-scene contacts is helpful for the estimation and synthesis [WXX+21, HCV+21] of plausible 3D human motions. Some existing works regress sparse joint contacts on a kinematic skeleton [LSC+19, SGX+21, SGXT20, RGH+20, RBH+21, ZYC+20] or sparse markers [ZZB+21]. A few approaches forecast contacts on a dense human mesh surface [HGT+21a, MOT+21]. Hassan *et al.* [HGT+21a] place a human in a 3D scene considering the semantic information and dense human body contact labels. Müller *et al.* [MOT+21] propose a dataset with discrete annotations for self-contacts on the human body. Consequently, they apply a self-contact loss for more plausible final 3D poses. Unlike the existing works, our algorithm estimates vertex-wise dense contact labels on the human body surface from an RGB input only. Along with that, it also regresses dense contact labels on the environment given the scene point cloud along with the RGB sequence. The simultaneous estimation of the body and scene contacts allows HULC to disambiguate the depth and scale of the subject, although only a single camera view and a single scene point cloud are used as inputs.

Monocular MoCap with scene interactions. Among the scene-aware MoCap approaches [SGX+21, SGXT20, RGH+20, HCTB19, ZZB+21, RBH+21, ZMS18], there are a few ones that consider human-environment interactions given a highly detailed scene geometry [HCTB19, ZZB+21, LSS+22]. PROX (PROX-D)[HCTB19] estimates 3D motions given RGB (RGB-D)

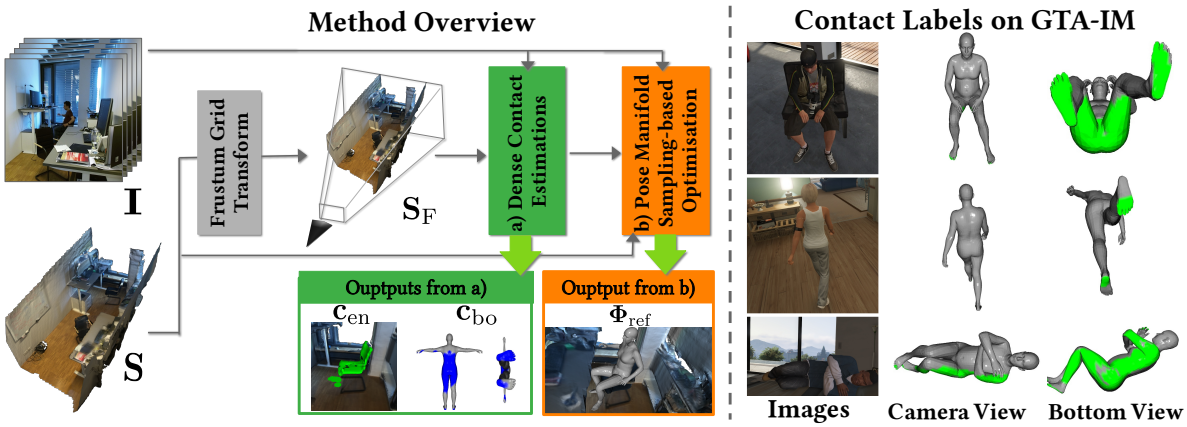


Figure 4.1: (*Left*) Given image sequence $\boldsymbol{I}$, scene point cloud $\boldsymbol{S}$ and its associated frustum voxel grid $\boldsymbol{S}_{\mathrm{F}}$, HULC first predicts for each frame dense contact labels on the body $\boldsymbol{c}_{\mathrm{bo}}$, and on the environment $\boldsymbol{c}_{\mathrm{en}}$. It then refines initial, physically-inaccurate and scale-ambiguous global 3D poses $\boldsymbol{\Phi}_0$ into the final ones $\boldsymbol{\Phi}_{\mathrm{ref}}$ in (b). Also see Fig. 4.2 for the details of stage (a) and (b). (*Right*) Example visualisations of our contact annotations (shown in green) on GTA-IM dataset [CGM+20].

image, along with an input geometry provided as a signed distance field (SDF). Given an RGB(D) measurement and a mesh of the environment, LEMO [ZZB+21] also produces geometry-aware global 3D human motions with an improved motion quality characterised by smoother transitions and robustness to occlusions thanks to the learned motion priors. These two algorithms require an RGB or RGB-D sequence with SDF (a 3D scan of the scene) or occlusion masks. In contrast, our HULC requires only an RGB image sequence and a point cloud of the scene; it returns dense contact labels on 1) the human body and 2) the environment, 3) global 3D human motion with translations and 4) absolute scale of the human body. See Table 4.1 for an overview of the characteristics. Compared to PROX and LEMO, HULC shows significantly-mitigated body-environment collisions.

Sampling-based human pose tracking. Several sampling-based human pose tracking algorithms were proposed. Some of them utilise particle-swarm optimisation [JTM10, SRSZ13, SRS+12]. Charles *et al.* [CPEZ13] employ Parzen windows for 2D joints tracking. Similar to our HULC, Sharma *et al.* [SVB+19] generate 3D pose samples by a conditional variational autoencoder (VAE) [SLY15] conditioned on 2D poses. In contrast, we utilise the learned pose manifold of VAE for sampling, which helps to avoid local minima and prevent body-scene collisions. Also, unlike [SVB+19], we sample around a latent vector obtained from the VAE's encoder to obtain poses that are plausible and similar to the input 3D pose.

4.3 Method

Given monocular video frames and a point cloud of the scene registered to the coordinate frame of the camera, our goal is to infer physically-plausible global 3D human poses along with dense contact labels on both body and environment surfaces. Our approach consists of two stages (Fig. 4.1):

- **Dense Body-environment contacts estimation**: Dense contact labels are predicted on body and scene surfaces using a learning-based approach with a pixel-aligned implicit representation inspired by [SHN+19] (Sec. 4.3.1);

- **Sampling-based optimisation on the pose manifold**: We combine sampling in a learned latent pose space with gradient descent to obtain the absolute scale of the subject and its global 3D pose, under hard guidance by predicted contacts. This approach significantly improves the accuracy of the estimated root translation and articulations, and mitigates incorrect environment penetrations. (Sec. 4.3.2).

Modelling and Notations. Our method takes as input a sequence $\boldsymbol{I} = \boldsymbol{I}_1, ..., \boldsymbol{I}_T$ of T successive video frames from a static camera with known intrinsics ($T = 5$ in our experiments). We detect a squared bounding box around the subject and resize the cropped image region to 225×225 pixels. The background scene's geometry that corresponds to the detected bounding box is represented by a single static point cloud $\boldsymbol{S} \in \mathbb{R}^{M\times 3}$ composed of M points aligned in the camera reference frame in an absolute scale. To model the 3D pose and human body surface, we employ the parametric model SMPL-X [PCG+19] (its gender-neutral version). This model defines the 3D body mesh as a differentiable function $\mathcal{M}(\boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\beta})$ of global root translation $\boldsymbol{\tau} \in \mathbb{R}^3$, global root orientation $\boldsymbol{\phi} \in \mathbb{R}^3$, root-relative pose $\boldsymbol{\theta} \in \mathbb{R}^{3K}$ of K joints and shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ capturing body's identity. For efficiency, we downsample the original SMPL-X body mesh with over 10k vertices to $\boldsymbol{V} \in \mathbb{R}^{N\times 3}$, where $N = 655$. In the following, we denote $\boldsymbol{V} = \mathcal{M}(\boldsymbol{\Phi}, \boldsymbol{\beta})$, where $\boldsymbol{\Phi} = (\boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{\theta})$ denotes the kinematic state of the human skeleton, from which the global positions $\boldsymbol{X} \in \mathbb{R}^{K\times 3}$ of the $K = 21$ joints can be derived.

4.3.1 Contact Estimation in the Scene

We now describe our learning-based approach for contact labels estimation on the human body and environment surfaces; see Fig. 4.1-a) for an overview of this stage. The approach takes $\boldsymbol{I}$ and $\boldsymbol{S}$ as inputs. It comprises three fully-convolutional feature extractors, N_1, N_2 and N_3, and two fully-connected layer-based contact prediction networks, Ω_{bo} and Ω_{en}, for body and environment, respectively.

Network N_1 extracts from $\boldsymbol{I}$ a stack of visual features $\text{f}_{\boldsymbol{I}} \in \mathbb{R}^{32\times 32\times 256}$. The latent space features of N_1 are also fed to Ω_{bo} to predict the vector $\boldsymbol{c}_{\text{bo}} \in [0, 1]^N$ of per-vertex contact probabilities on the *body* surface.

We also aim at estimating the corresponding contacts on the *environment* surface using an implicit function. To train a model that generalises well, we need to address two challenges: (i) No correspondence information between the scene points and the image pixels are given; (ii) Each scene contains a variable number of points. Accordingly, we convert the scene point cloud $\boldsymbol{S}$ into a frustum voxel grid $\boldsymbol{S}_{\text{F}} \in \mathbb{R}^{32\times 32\times 256}$ (the third dimension corresponds to the discretised depth of the 3D space over 256 bins, please refer to our supplement for the details). This new representation is independent of the original point-cloud size and is aligned with the camera's view direction. The latter will allow us to leverage a pixel-aligned implicit function inspired by PIFu [SHN+19], which helps the networks figure out the correspondences between pixel and geometry information. More specifically, $\boldsymbol{S}_{\text{F}}$ is fed into N_2, which returns scene features $\text{f}_{\text{S}} \in \mathbb{R}^{32\times 32\times 256}$. The third encoder, N_3, ingests f_{I} and f_{S} concatenated along their third dimension and returns pixel-aligned features $\text{F}_{\text{P}} \in \mathbb{R}^{32\times 32\times 64}$. Based on F_{P}, Ω_{en} predicts the contact labels on the environment surface as follows. Given a 3D position in the scene, we extract the corresponding visual feature $\text{f}_{\text{P}} \in \mathbb{R}^{64}$ at the (u, v)-position in the image space from F_{P} (via spacial bilinear interpolation), and query arbitrary depth with a one-hot vector $\text{f}_{\text{z}} \in \mathbb{R}^{256}$. We next estimate the contact labels c_{en} as follows:

$$c_{\text{en}} = \Omega_{\text{en}}(\text{f}_{\text{P}}, \text{f}_{\text{z}}). \tag{4.1}$$

Given contact ground truths $\hat{\boldsymbol{c}}_{\text{bo}} \in \{0, 1\}^N$ and $\hat{\boldsymbol{c}}_{\text{en}} \in \{0, 1\}^M$ on the body and the environment,

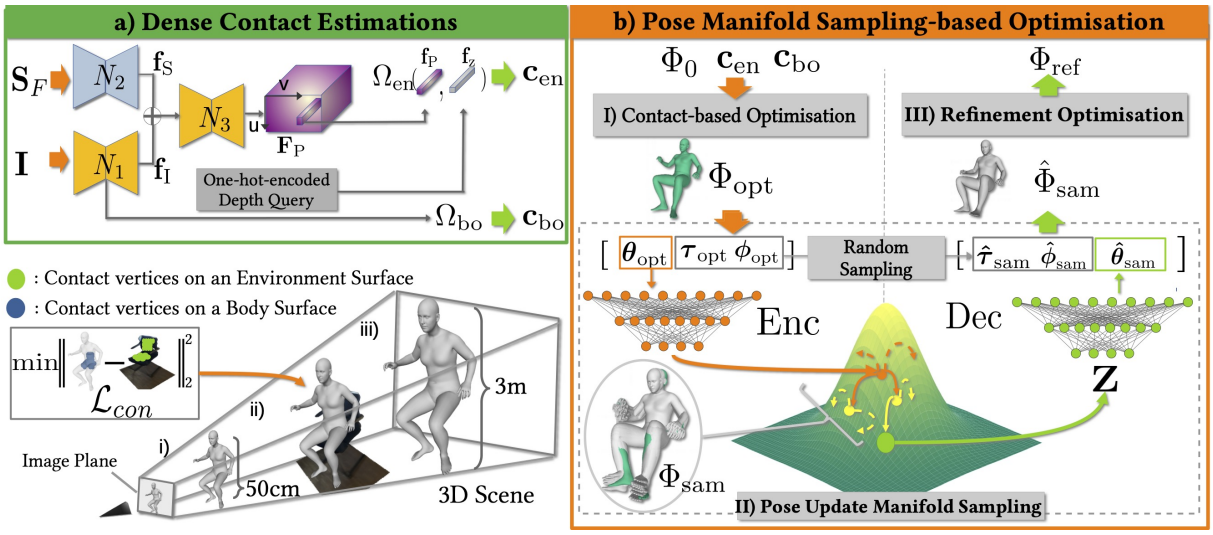


Figure 4.2: **Overview of a) dense contact estimation and b) pose manifold sampling-based optimisation.** In b-II), we first generate samples around the mapping from $\boldsymbol{\theta}_{\text{opt}}$ (orange arrows), and elite samples are then selected among them (yellow points). After resampling around the elite samples (yellow arrows), the best sample is selected (green point). The generated sample poses $\boldsymbol{Phi}_{\text{sam}}$ (in gray colour at the bottom left in b-II)) from the sampled latent vectors are plausible and similar to $\boldsymbol{Phi}_{\text{opt}}$. **(*bottom left of the Figure*)** Different body scale and depth combinations can be re-projected to the same image coordinates (i, ii and iii), *i.e.*, **scale-depth ambiguity**. To simultaneously estimate the accurate body scale and depth of the subject (ii), we combine the body-environment contact surface distance loss $\mathcal{L}_{\text{con}}$ with the 2D reprojection loss.

the five networks are trained with the following loss:

$$\mathcal{L}_{\text{labels}} = \|\boldsymbol{c}_{\text{en}} - \hat{\boldsymbol{c}}_{\text{en}}\|_2^2 + \lambda \, \text{BCE}(\boldsymbol{c}_{\text{bo}}, \hat{\boldsymbol{c}}_{\text{bo}}), \tag{4.2}$$

where BCE denotes the binary cross-entropy and $\lambda = 0.3$. We use BCE for the body because the ground-truth contacts on its surface are binary; the ℓ_2 loss is used for the environment, as sparse ground-truth contact labels are smoothed with a Gaussian kernel to obtain continuous signals. For further discussions of (4.2), please refer to our supplement. At test time, we only provide the 3D vertex positions of the environment to $\Omega_{\text{en}}(\cdot)$—to find the contact area on the scene point cloud—rather than all possible 3D sampling points as queries. This significantly accelerates the search of environmental contact labels while reducing the number of false-positive contact classifications. For more details of the network architecture, further discussions of the design choice and data pre-processing, please refer to our supplement.

4.3.2 Pose Manifold Sampling-based Optimisation

In the second stage of the approach, we aim at recovering an accurate global 3D trajectory of the subject as observed in the video sequence, see Fig. 4.2-(b) for the overview. An initial estimate $\boldsymbol{\Phi}_0$ is extracted for each input image using SMPLify-X [PCG+19]. Its root translation $\boldsymbol{\tau}$ being subject to scale ambiguity, we propose to estimate it more accurately, along with the actual scale h of the person with respect to the original body model's height, under the guidance of the predicted body-environment contacts (**Contact-based Optimisation**).We then update the body trajectory and articulations in the scene, while mitigating the body-environment collisions

with a new sampling-based optimisation on the pose manifold (**Sampling-based Trajectory Optimisation**). A subsequent refinement step yields the final global physically-plausible 3D motions.

I) Contact-based Optimisation Scale ambiguity is inherent to a monocular MoCap setting: Human bodies with different scale and depth combinations in 3D can be reprojected on the same positions in the image frame; see Fig. 4.2 and supplementary video for the schematic visualisation. Most existing algorithms that estimate global 3D translations of a subject either assume its known body scale [SGXT20, DSJ+21, SGX+21] or use a statistical average body scale [MSS+17]. In the latter case, the estimated $\boldsymbol{tau}$ is often inaccurate and causes physically implausible body-environment penetrations. In contrast to the prior art, we simultaneously estimate $\boldsymbol{tau}$ and h by making use of the body-environment dense contact labels from the previous stage (Sec. 4.3.1).

For the given frame at time $t \in [\![1, T]\!]$, we select the surface regions with $\boldsymbol{c}_{\text{en}} > 0.5$ and $\boldsymbol{c}_{\text{bo}} > 0.5$ as effective contacts and leverage them in our optimisation. Let us denote the corresponding index subsets of body vertices and scene points by $\mathcal{C}_{\text{bo}} \subset [\![1, N]\!]$ and $\mathcal{C}_{\text{en}} \subset [\![1, M]\!]$. The objective function for contact-based optimisation is defined as:

$$\mathcal{L}_{\text{opt}}(\boldsymbol{\tau}, h) = \lambda_{\text{2D}}\mathcal{L}_{\text{2D}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}, \tag{4.3}$$

where the reprojection $\mathcal{L}_{\text{2D}}$, the temporal smoothness $\mathcal{L}_{\text{smooth}}$ and the contact $\mathcal{L}_{\text{con}}$ losses weighted by empirically-set multipliers λ_{2D}, λ_{smooth} and λ_{con}, read:

$$\mathcal{L}_{\text{2D}} = \frac{1}{K}\sum_{k=1}^{K} w_k \left\| \Pi(\boldsymbol{X}_k) - \boldsymbol{p}_k \right\|_2^2, \tag{4.4}$$

$$\mathcal{L}_{\text{smooth}} = \left\| \boldsymbol{\tau} - \boldsymbol{\tau}_{\text{prev}} \right\|_2^2, \tag{4.5}$$

$$\mathcal{L}_{\text{con}} = \sum_{n \in \mathcal{C}_{\text{bo}}} \min_{m \in \mathcal{C}_{\text{en}}} \left\| \boldsymbol{V}_n - \boldsymbol{P}_m \right\|_2^2, \tag{4.6}$$

where $\boldsymbol{p}_k$ and w_k are the 2D detection in the image of the k-th body joint and its associated confidence, respectively, obtained by OpenPose [CHS+19]; Π is the perspective projection operator; $\boldsymbol{\tau}_{\text{prev}}$ is the root translation estimated in the previous frame; $\boldsymbol{X}_k$, $\boldsymbol{V}_n$ and $\boldsymbol{P}_m$ are, respectively, the k-th 3D joint, the n-th body vertex ($n \in \mathcal{C}_{\text{bo}}$) and the m-th scene point ($m \in \mathcal{C}_{\text{en}}$). Note that the relative rotation and pose are taken from $\boldsymbol{\Phi}_0$. The body joints and vertices are obtained from $\mathcal{M}$ using $\boldsymbol{tau}$ and scaled with h. For $\mathcal{L}_{\text{con}}$, we use a directed Hausdorff measure [KLSW09] as a distance between the body and environment contact surfaces. The combination of $\mathcal{L}_{\text{con}}$ and $\mathcal{L}_{\text{2D}}$ is key to disambiguate $\boldsymbol{\tau}$ and h (thus, resolving the monocular scale ambiguity). As a result of optimising (4.3) in frame t, we obtain $\boldsymbol{\Phi}^t_{\text{opt}}$, *i.e.,* the global 3D human motion with absolute body scale. We solve jointly on T frames and optimise for a single h for them.

II-a) Sampling-based Trajectory Optimisation Although the poses $\boldsymbol{\Phi}^t_{\text{opt}}$, $t = 1 \cdots T$, estimated in the previous step yield much more accurate $\boldsymbol{\tau}$ and h compared to existing monocular RGB-based methods, incorrect body-environment penetrations are still observable. This is because the gradient-based optimisation often gets stuck in bad local minima (see the supplementary video for a toy example illustrating this issue). To overcome this problem, we introduce an additional sampling-based optimisation that imposes hard penetration constraints, thus significantly mitigating physically-implausible collisions. The overview of this algorithm is as follows: (i) For each frame t, we first draw candidate poses around $\boldsymbol{\Phi}^t_{\text{opt}}$ with a sampling function $\mathcal{G}$; (ii) The quality of these samples is ranked by a function $\mathcal{E}$ that allows selecting the most promising ("elite") ones; samples with severe collisions are discarded; (iii) Using $\mathcal{G}$ and $\mathcal{E}$ again,

we generate and select new samples around the elite ones. The details of these steps, $\mathcal{E}$ and $\mathcal{G}$, are elaborated next (dropping time index t for simplicity).

II-b) Generating Pose Samples. We aim to generate N_{sam} sample states $\boldsymbol{\Phi}_{\text{sam}}$ around the previously-estimated $\boldsymbol{\Phi}_{\text{opt}} = (\boldsymbol{\tau}_{\text{opt}}, \boldsymbol{\phi}_{\text{opt}}, \boldsymbol{\theta}_{\text{opt}})$. To generate samples $(\boldsymbol{\tau}_{\text{sam}}, \boldsymbol{\phi}_{\text{sam}})$ for the global translation and orientation, with 3DoF each, we simply use a uniform distribution around $(\boldsymbol{\tau}_{\text{opt}}, \boldsymbol{\phi}_{\text{opt}})$; see our supplement for the details. However, naïvely generating the relative pose $\boldsymbol{\theta}_{\text{sam}}$ in the same way around $\boldsymbol{\theta}_{\text{opt}}$ is highly inefficient because (i) the body pose is high-dimensional and (ii) the randomly-sampled poses are not necessarily plausible. These reasons lead to an infeasible amount of generated samples required to find a plausible collision-free pose; which is intractable on standard graphics hardware. To tackle these issues, we resort to the pose manifold learned by VPoser [PCG$^+$19], which is a VAE [KW14] trained on AMASS [MGT$^+$19], *i.e.,* a dataset with many highly accurate MoCap sequences. Sampling is conducted in this VAE's latent space rather than in the kinematics pose space. Specifically, we first map $\boldsymbol{\theta}_{\text{opt}}$ into a latent pose vector with the VAE's encoder $\text{Enc}(\cdot)$. Next, we sample latent vectors using a Gaussian distribution centered at this vector, with standard deviation $\boldsymbol{\sigma}$ (see Fig. 4.2-b). Each latent sample is then mapped through VAE's decoder $\text{Dec}(\cdot)$ into a pose that is combined with the original one on a per-joint basis. The complete sampling process reads:

$$\boldsymbol{Z} \sim \mathcal{N}(\,\text{Enc}(\boldsymbol{\theta}_{\text{opt}}), \boldsymbol{\sigma}), \;\; \boldsymbol{\theta}_{\text{sam}} = \boldsymbol{w} \circ \boldsymbol{\theta}_{\text{opt}} + (1 - \boldsymbol{w}) \circ \text{Dec}(\boldsymbol{Z}), \tag{4.7}$$

where $\circ$ denotes Hadamard matrix product and $\boldsymbol{w} \in \mathbb{R}^{3K}$ is composed of the detection confidence values w_k, $k = 1 \cdots K$, obtained from OpenPose, each appearing three times (for each DoF of the joint). This confidence-based strategy allows weighting higher the joint angles obtained by sampling, if the image-based detections are less confident (*e.g.,* under occlusions). Conversely, significant modifications are not required for the joints with high confidence values.

Since the manifold learned by VAE is smooth, the poses derived from the latent vectors sampled around $\text{Enc}(\boldsymbol{\theta}_{\text{opt}})$ should be close to $\boldsymbol{\theta}_{\text{opt}}$. Therefore, we empirically set $\boldsymbol{\sigma}$ to a small value (0.1). Compared to the naïve random sampling in the joint angle space, whose generated poses are not necessarily plausible, this pose sampling on the learned manifold significantly narrows down the solution space. Hence, a lot fewer samples are required to escape local minima. At the bottom left of Fig. 4.2-b contains examples (gray colour) of $\boldsymbol{\Phi}_{\text{sam}}$ ($N_{\text{sam}} = 10$) overlayed onto $\boldsymbol{\Phi}_{\text{opt}}$ (green). In the following, we refer to this sample generation process as function $\mathcal{G}(\cdot)$.

II-c) Sample Selection. The quality of the N_{sam} generated samples $\boldsymbol{\Phi}_{\text{sam}}$ is evaluated using the following cost function:

$$\mathcal{L}_{\text{sam}} = \mathcal{L}_{\text{opt}} + \lambda_{\text{sli}}\mathcal{L}_{\text{sli}} + \lambda_{\text{data}}\mathcal{L}_{\text{data}}, \tag{4.8}$$

$$\mathcal{L}_{\text{sli}} = \|\boldsymbol{V}_{\text{c}} - \boldsymbol{V}_{\text{c,pre}}\|_2^2, \tag{4.9}$$

$$\mathcal{L}_{\text{data}} = \|\boldsymbol{\Phi}_{\text{sam}} - \boldsymbol{\Phi}_{\text{opt}}\|_2^2, \tag{4.10}$$

where $\mathcal{L}_{\text{sli}}$ and $\mathcal{L}_{\text{data}}$ are contact sliding loss and data loss, respectively, and $\mathcal{L}_{\text{opt}}$ is the same as in (4.3) with the modification that the temporal consistency (4.5) applies to the whole $\boldsymbol{\Phi}_{\text{sam}}$; $\boldsymbol{V}_{\text{c}}$ and $\boldsymbol{V}_{\text{c,pre}}$ are the body contact vertices (with vertex indices in $\mathcal{C}_{\text{bo}}$) and their previous positions, respectively.

Among N_{sam} samples ordered according to their increasing $\mathcal{L}_{\text{sam}}$ values, the selection function $\mathcal{E}_U(\cdot)$ first discards those causing stronger penetrations (in the sense that the amount of scene points inside a human body is above a threshold γ) and returns U first samples from the remaining ones. If no samples pass the collision test, we regenerate the new set of N_{sam} samples. This selection mechanism introduces the collision handling in a hard manner. After applying $\mathcal{E}_U(\cdot)$, with $U < N_{\text{sam}}$, U elite samples are retained. Then, $\lfloor N_{\text{sam}}/U \rfloor$ new samples are regenerated

around every elite sample using $\mathcal{G}$. Among those, the one with minimum $\mathcal{L}_{\text{sam}}$ value is retained as the final estimate. The sequence of obtained poses is temporally smoothed by Gaussian filtering to further remove jittering, which yields the global 3D motion $(\hat{\boldsymbol{P}\boldsymbol{h}i}^{t}_{\text{sam}})_{t=1}^{T}$ with significantly mitigated collisions.

III) Final Refinement. In previous steps we obtained the sequence $\hat{\boldsymbol{\Phi}}_{\text{sam}} = (\hat{\boldsymbol{\tau}}_{\text{sam}}, \hat{\boldsymbol{\phi}}_{\text{sam}}, \hat{\boldsymbol{\theta}}_{\text{sam}})$ of kinematic states whose severe body-environment collisions are prevented as hard constraints. Starting from these states as initialisation, we perform a final gradient-based refinement using cost function $\mathcal{L}_{\text{sam}}$ with $\hat{\boldsymbol{\Phi}}_{\text{sam}}$ replacing $\boldsymbol{\Phi}_{\text{opt}}$. The final sequence is denoted $(\boldsymbol{\Phi}^{t}_{\text{ref}})_{t=1}^{T}$.

4.4 Datasets with Contact Annotations

As there are no publicly-available large-scale datasets with images and corresponding human-scene contact annotations, we annotate several existing datasets.

GTA-IM [CGM+20] dataset contains various daily 3D motions. First, we fit SMPL-X model onto the 3D joint trajectories in GTA-IM. For each frame, we select contact vertices on the human mesh if: i) The Euclidean distance between the human body vertices on and the scene vertices are smaller than a certain threshold; ii) The velocity of the vertex is lower than a certain threshold. In total, we obtain the body surface contact annotations on $320k$ frames, which will be released for research purposes, see Fig. 4.1 for the examples of the annotated contact labels.

PROX dataset [HCTB19] contains scanned scene meshes, scene SDFs, RGB-D sequences, 3D human poses and shapes generated by fitting SMPL-X model onto the RGB-D sequences (considering collisions). We consider the body vertices, whose SDF values are lower than 5 cm, as contacts. We annotate the environment contacts by finding the vertices that are the nearest to the body contacts.

GPA dataset [WCR+22, WSF20] contains multi-view image sequences of people interacting with various rigid 3D geometries, accurately reconstructed 3D scenes and 3D human motions obtained from VICON system [Vic] with 28 calibrated cameras. We fit SMPL-X on GPA to obtain the 3D shapes and compute the scene's SDFs to run other methods [HCTB19, ZZB+21, HGT+21a].

We extract from **GPA** 14 test sequences with 5 different subjects. We also split **PROX** [HCTB19] into training and test sequences. The training sequences of **PROX** and **GTA-IM** [CGM+20] are used to train the contact estimation networks. For further details of dataset and training, please refer to our supplement.

4.5 Evaluations

We compare our HULC with the most related scene-aware 3D MoCap algorithms, *i.e.,* PROX [HCTB19], PROX-D [HCTB19], POSA [HGT+21a] and LEMO [ZZB+21]. We also test SMPLify-X [PCG+19] which does not use scene constraints. The root translation of SMPLify-X is obtained from its estimated camera poses as done in [HCTB19]. To run LEMO [ZZB+21] on the RGB sequence, we use SMPLify-X[PCG+19] to initialise it; we call this combination "LEMO (RGB)".

We use the selected test sequences of GPA [WCR+22, WSF20] and PROX [HCTB19] dataset for the quantitative and qualitative comparisons. To avoid redundancy, we downsample all the predictions to 10 fps except for the temporal consistency measurement (e_{smooth} in Table 4.4). Since the 3D poses in PROX dataset are prone to inaccuracies due to their human model fitting onto the RGB-D sequence, we use it only for reporting the body-scene penetrations (Table 4.4) and for qualitative comparisons.

Table 4.2: Comparisons of 3D error on GPA dataset [WCR+22, WSF20]. "†" denotes that the occlusion masks for LEMO(RGB) were computed from GT 3D human mesh.

	No Procrustes			Procrustes		
	MPJPE [mm]↓	PCK [%]↑	PVE [mm]↓	MPJPE [mm]↓	PCK [%]↑	PVE [mm]↓
Ours	**217.9**	**35.3**	**214.7**	**81.5**	**89.3**	**72.6**
Ours (w/o S)	221.3	34.5	217.2	82.6	**89.3**	73.1
Ours (w/o R)	240.8	31.9	237.3	83.1	86.6	73.6
Ours (w/o SR)	251.1	31.5	245.2	83.9	86.6	74.1
SMPLify-X [PCG+19]	550.0	10.0	549.1	84.7	85.9	74.1
PROX [HCTB19]	549.7	10.1	548.7	84.6	86.0	73.9
POSA [HGT+21a]	552.2	10.1	550.9	85.5	85.6	74.5
LEMO (RGB) [ZZB+21]	570.1	8.75	570.5	83.0	86.4	73.7
LEMO (RGB) [ZZB+21]†	570.0	8.77	570.4	83.0	86.4	73.6

Table 4.3: Ablations and comparisons for global translations and absolute body length on GPA dataset.

	global translation error [m] ↓	absolute bone length error [m]↓
Ours (+1m)	**0.242**	0.104
Ours (+3m)	0.244	**0.097**
Ours (+10m)	0.244	0.109
Baseline (+1m)	0.751	0.498
Baseline (+3m)	1.033	0.560
Baseline (+10m)	2.861	1.918
SMPLify-X [PCG+19]	0.527	0.156
PROX [HCTB19]	0.528	0.160
POSA [HGT+21a]	0.545	0.136

4.5.1 Quantitative Results

We report 3D joint and vertex errors (Table 4.2), global translation and body scale estimation errors (Table 4.3), body-environment penetration and smoothness errors (Table 4.4) and ablations on the sampling-based optimisation component, *i.e.*, a) Manifold sampling vs. random sampling and b) Different number of sampling iterations in Fig. 4.3. "Ours (w/o S)" represents our method without the sampling optimisation component, *i.e.,* only the contact-based optimisation and refinement are applied (see Fig. 4.2-(b) and Sec. 4.3.2). "Ours (w/o R)" represents our method without the final refinement. "Ours (w/o SR)" denotes ours without the sampling and refinement. For a further ablation study and evaluation of contact label estimation networks, please see our supplement.

3D Joint and Vertex Errors. Table 4.2 compares the accuracy of 3D joint and vertex positions with and without Procrustes alignment. LEMO also requires human body occlusion masks on each frame. We compute them using the scene geometry and SMPLify-X [PCG+19] results. We also show another variant "LEMO (RGB)†" whose occlusion masks are computed using the ground-truth global 3D human mesh instead of SMPLify-X. Here, we report the standard 3D metrics, *i.e.,* mean per joint position error (MPJPE), percentage of correct keypoints (PCK) (@150mm) and mean per vertex error (PVE). Lower MPJPE and PVE represent more accurate 3D reconstructions, higher PCK indicates more accurate 3D joint positions.

On all these metrics, HULC outperforms other methods both with and without Procrustes.

Table 4.4: Comparisons of physical plausibility measures on GPA dataset [WCR+22, WSF20] and PROX dataset [HCTB19].

		GPA Dataset		PROX Dataset
		non penet. [%]↑	e_{smooth} ↓	non penet. [%]↑
RGB	Ours	**99.4**	20.2	97.0
	Ours (w/o S)	97.6	28.1	93.8
	Ours (w/o R)	**99.4**	24.7	**97.1**
	Ours (w/o SR)	97.6	47.1	93.8
	SMPLify-X [PCG+19]	97.7	43.3	88.9
	PROX [HCTB19]	97.7	43.2	89.8
	LEMO (RGB)[ZZB+21]	97.8	**19.9**	-
	POSA [HGT+21a]	98.0	47.0	93.0
RGB-D	PROX-D [HCTB19]	-	-	94.2
	LEMO [ZZB+21]	-	-	96.4

Notably, thanks to substantially more accurate global translations obtained from the contact-based optimisation (Sec. 4.3.2), HULC significantly reduces the MPJPE and PVE with a big margin, *i.e.,* ≈60% error deduction in MPJPE and PVE w/o Procrustes compared to the second-best method. The ablative studies on Table 4.2 also indicate that both the sampling and refinement optimisations contribute to accurate 3D poses. Note that the sampling optimisation alone ("Ours (w/o R)") does not significantly reduce the error compared to "Ours (w/o SR)". This is because the sampling component prioritises removal of environment penetrations by introducing hard collision handling, which is the most important feature of this component. Therefore, the sampling component significantly contributes to reducing the environment collision as can be seen in Table 4.4 (discussed in the later paragraph). Applying the refinement after escaping from severe penetrations by the sampling optimisation further increases the 3D accuracy ("Ours" in Table 4.2) while significantly mitigating physically implausible body-environment penetrations (Table 4.4).

Global Translation and Body Scale Estimation. Table 4.3 reports global translation and body scale estimation errors for the ablation study of the contact-based optimisation (Sec. 4.3.2). More specifically, we evaluate the output Φ_{opt} obtained from the contact-based optimisation denoted "ours". We also show the optimisation result without using the contact loss term (4.6) ("Baseline"). The numbers next to the method names represent the initialisation offset from the ground-truth 3D translation position (*e.g.,* "+10m" indicates that the initial root position of the human body was placed at 10 meters away along the depth direction from the ground-truth root position when solving the optimisations).

Without the contact loss term—since global translation and body scale are jointly estimated in the optimisation—the baseline method suffers from *up-to-scale* issue (see Fig. 4.2). Hence, its results are significantly worse due to worse initialisations. In contrast, our contact-based optimisation disambiguates the scale and depth by localising the contact positions on the environment, which confirms HULC to be highly robust to bad initialisations. Compared to the RGB-based methods PROX, POSA and SMPLify-X, our contact-based optimisation result has ≈40% smaller error in the absolute bone length, and ≈57% smaller error in global translation, which also contributes to the reduced body-environment collisions as demonstrated in Table 4.4 (discussed in the next paragraph).

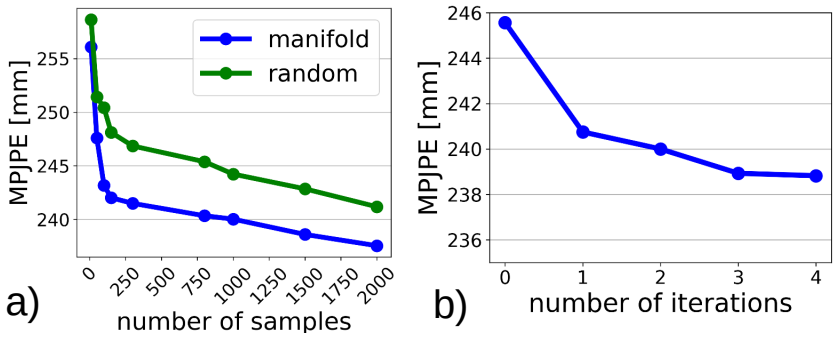


Figure 4.3: (a) MPJPE [mm] comparison with different numbers of samples for the learned manifold sampling strategy vs. the naïve random sampling in the joint angle space of the kinematic skeleton.(b) MPJPE [mm] comparison with different numbers of iterations in the sampling strategy.

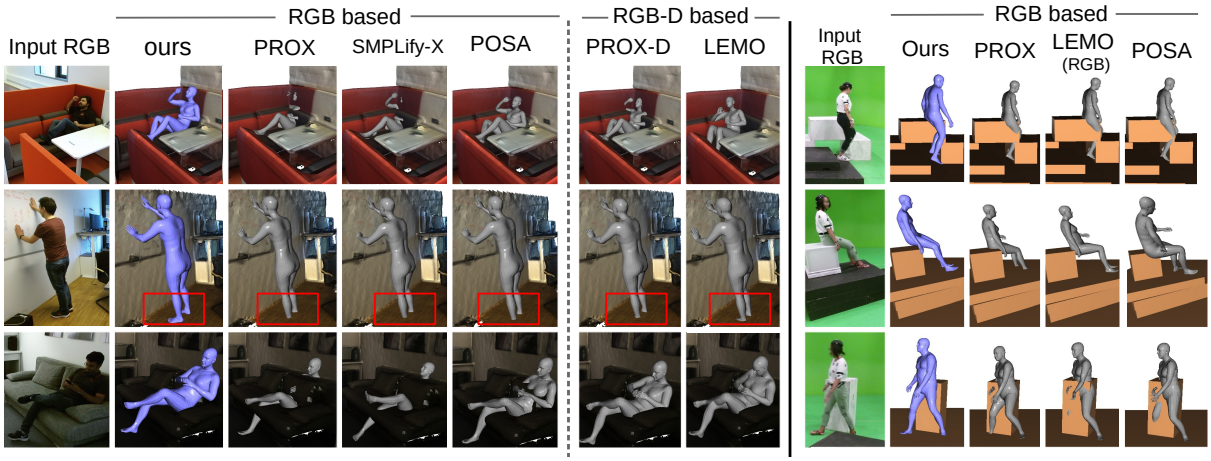


Figure 4.4: The qualitative comparisons of our results with the related methods on PROX (left) and GPA dataset (right). Our RGB-based HULC shows fewer body-scene penetrations even when compared with RGB-D based methods; mind the red rectangles in the second row.

Plausibility Measurements. We also report the plausibility of the reconstructed 3D motions in Table 4.4. *Non penet.* measures the average ratio of non-penetrating body vertices into the environment over all frames. A higher value denotes fewer body-environment collisions in the sequence; e_{smooth} measures the temporal smoothness error proposed in [SGXT20]. Lower e_{smooth} indicates more temporally smooth 3D motions. On both GPA and PROX datasets, our full framework mitigates the collisions thanks to the manifold sampling-based optimisations (ours vs. ours (w/o S)). It also does so when compared to other related works as well. Notably, HULC shows the least amount of collisions even compared with RGBD-based methods on the PROX dataset. Finally, the proposed method also shows the significantly low e_{smooth} (on par with LEMO(RGB)) in this experiment.

More Ablations on Sampling-based Optimisation. In addition to the ablation studies reported in Tables 4.2, 4.3 and 4.4, we further assess the performance of the pose update manifold sampling step (Fig. 4.2-(b)-(II)) on GPA dataset [WCR$^+$22, WSF20], reporting the 3D error (MPJPE [mm]) measured in world frame. Note that we report MPJPE without the final

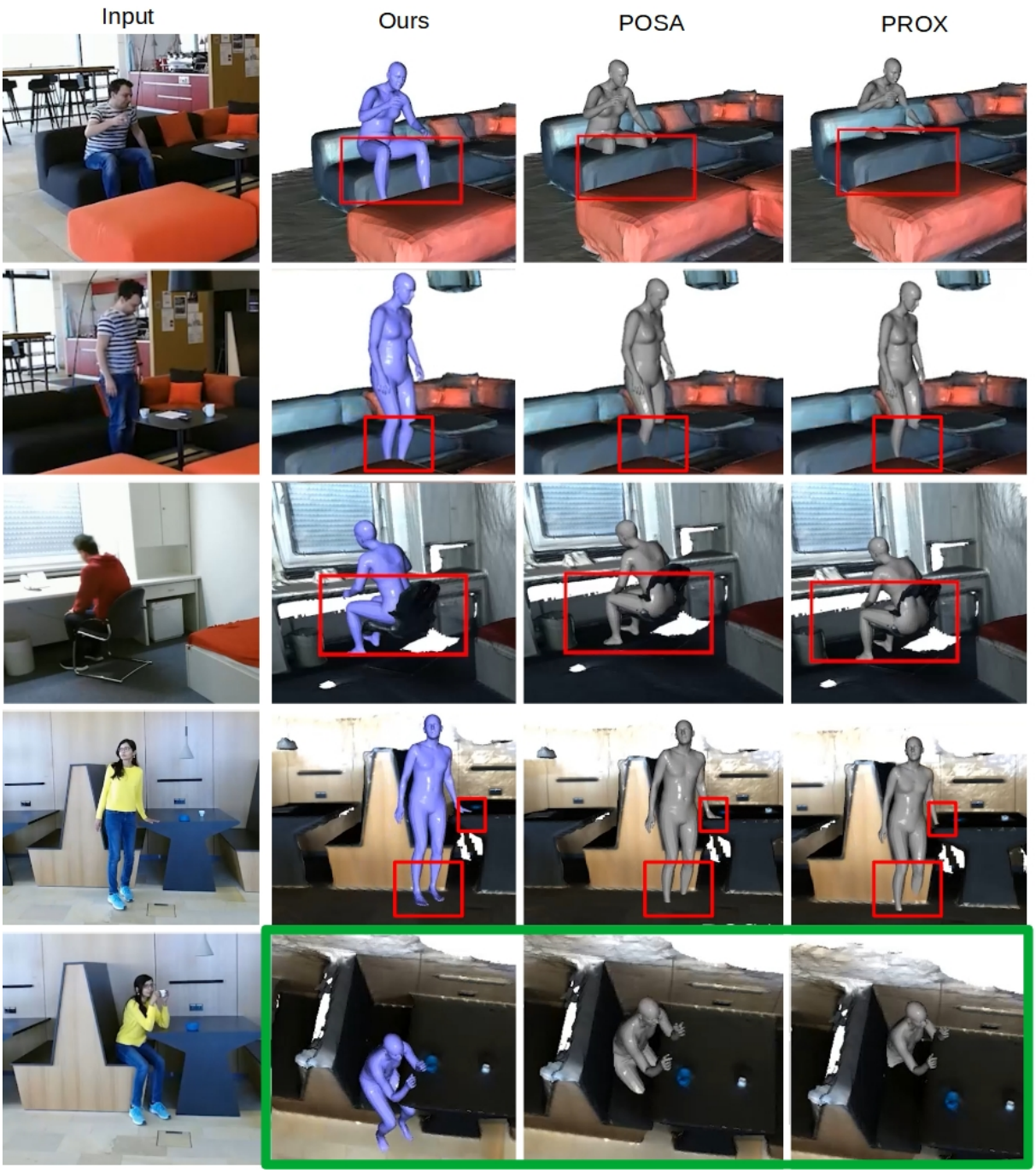


Figure 4.5: Qualitative comparison of our HULC *vs* existing scene-aware RGB-based methods on PROX [HCTB19] dataset. Our method shows significantly mitigated collisions thanks to our novel sampling-based optimisation, which handles the severe body-environment penetrations in a hard manner (red rectangles). We also show the results from a top view (green rectangle). Thanks to the contact-based optimisation using the estimated dense contacts on the body surfaces and the environment, our estimated 3D global root positions are significantly more accurate compared to the previous methods.

refinement step to assess the importance of the manifold sampling approach. In Fig. 4.3-(a), we show the influence of the number N_{sam} of samples on the performance of our manifold sampling

strategy vs. a naïve random sampling with a uniform distribution in a kinematic skeleton frame. For the details of the naïve random sampling strategy, please refer to our supplement. In Fig. 4.3-(a), since the generated samples of the learned manifold return plausible pose samples, our pose manifold sampling strategy requires significantly fewer samples compared to the random sampling (∼15× more samples are required for the random sampling to reach 243 [mm] error in MPJPE). This result strongly supports the importance of the learned manifold sampling. No more than 2000 samples can be generated due to the hardware memory capacity. In Fig. 4.3-(b), we report the influence of the number of generation-selection steps using functions $\mathcal{G}$ and $\mathcal{E}_U$ (with $U = 3$) introduced in Section 4.3.2, with $N_{\text{sam}} = 1000$ samples. No iteration stands for choosing the best sample from the first generated batch (hence no resampling), while one iteration is the variant described in Sec. 4.3.2. This first iteration sharply reduces the MPJPE, while the benefit of the additional iterations is less pronounced. Based on these observations, we use only one re-sampling iteration with 1000 samples in the previous experiments. Finally, we ablate the confidence value-based pose merging in Eq. (4.7), setting N_{sam}=1000 and the number of iterations to 0. The measured MPJPE for with and without this confidence merging are 245.5 and 249.1, respectively.

4.5.2 Qualitative Results

Figure 4.4 summarises the qualitative comparisons on GPA and PROX datasets. HULC produces more physically-plausible global 3D poses with mitigated collisions, whereas the other methods show body-environment penetrations. Even compared with the RGB(D) approaches, HULC mitigates collisions (mind the red rectangles). In Fig. 4.5, we further show comparisons of our method with other RGB-based scene-aware methods POSA[HGT$^+$21a] and PROX [HCTB19]. Our method shows physically more plausible interactions with the environment than the others. We also visualise the result from a bird's eye view to show the significance of the contact-based optimisation, which contributes to substantially more accurate global translation estimation than other related methods (green rectangle).

4.6 Conclusion

Limitations. HULC requires the scene geometry aligned in a camera frame like other related works [HCTB19, ZZB$^+$21, HGT$^+$21a]. Also, HULC does not capture non-rigid deformations of scenes and bodies, although the body surface and some objects in the environment deform (*e.g.,* when sitting on a couch or lying in a bed). Moreover, since our algorithm relies on the initial root-relative pose obtained from an RGB-based MoCap algorithm, the subsequent steps can fail under severe occlusions. Although the estimated contact labels help to significantly reduce the 3D translation error, the estimated environment contacts contain observable false positives. These limitations can be tackled in the future.

Conclusion. We introduced *HULC*—the first RGB-based scene-aware MoCap algorithm that estimates and is guided by dense body-environment surface contact labels combined with a pose manifold sampling. HULC shows 60% smaller 3D-localisation errors compared to the previous methods. Furthermore, deep body-environment collisions are handled in hard manner in the pose manifold sampling-based optimisation, which significantly mitigates collisions with the scene. HULC shows the lowest collisions even compared with RGBD-based scene-aware methods.

5 Human Motion Capture with Scene Deformation Recovery

Contents

3D human motion capture from monocular RGB images respecting interactions of a subject with complex and possibly deformable environments is a very challenging, ill-posed and under-explored problem. Existing methods address it only weakly and do not model possible surface deformations often occurring when humans interact with scene surfaces. In contrast, this paper proposes MoCapDeform, *i.e.,* a new framework for monocular 3D human motion capture that is the first to explicitly model non-rigid deformations of a 3D scene for improved 3D human pose estimation and deformable environment reconstruction. MoCapDeform accepts a monocular RGB video and a 3D scene mesh aligned in the camera space. It first localises a subject in the input monocular video along with dense contact labels using a new raycasting based strategy. Next, our human-environment interaction constraints are leveraged to jointly optimise global 3D human poses and non-rigid surface deformations. MoCapDeform achieves superior accuracy than competing methods on several datasets, including our newly recorded one with deforming background scenes.

This chapter is based on [LSS+22]. As the first author, Zhi Li worked on all parts of the project, including idea proposal, all implementation and experiments, and paper writing.

5.1 Introduction

3D human motion capture from monocular images is an active research area [MHRL17, PZZD18, TRA17, MSS+17, RSF18, KBJM18, DGM+19, SGXT20, DSJ+21, WLR22, AWS+22]. Relying solely on monocular RGB inputs is challenging and severely ill-posed, as no explicit 3D cues are provided. As observed in daily life and noticed in the literature [Gib50, HCTB19], the 3D world constrains human actions when they move and interact with it. Thus, environmental constraints and scene priors can provide additional cues for global 3D human motion capture. Leveraging them is a promising route to reduce the ambiguities and, at the same time, infer deformable scene geometry by observing human-scene interactions.

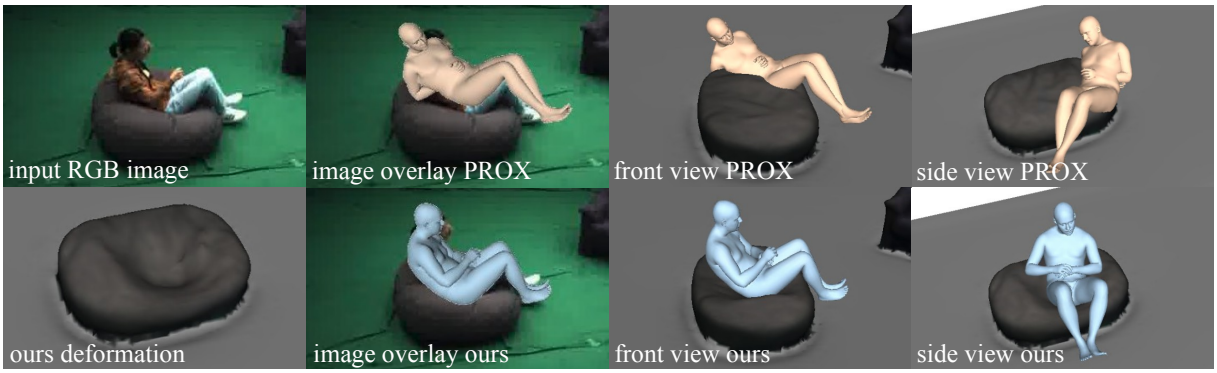


Figure 5.1: Existing monocular 3D human motion capture methods such as PROX [HCTB19] ignore abundant scene deformation when penalising human-scene collisions, resulting in erroneous global poses (top). **Our MoCapDeform algorithm is the first that models non-rigid scene deformations and finds the accurate global 3D poses of the subject by human-deformable scene interaction constraints**, achieving increased accuracy with significantly fewer penetrations (bottom).

Existing works consider foot-floor interactions to recover physically plausible human motions [SGXT20, RGH+20, SGX+21, RBH+21]. Others utilise explicit 3D environment models as constraints [HCTB19, ZPJ+20, WY21, ZZB+21, GMSPM21]. These techniques either do not fully address the scale ambiguity [SGXT20, RGH+20, SGX+21, RBH+21, HCTB19, ZPJ+20, WY21] or require RGB-D rather than RGB inputs [ZZB+21]. Several other methods rely on body-mounted sensors (inertial measurement units) [GMSPM21, YZH+22] to localise humans globally. Next, virtually all approaches consider the background being static and ignore potential scene changes caused by human-scene interactions. When a subject sits on a couch (lays in a bed), the latter deforms significantly due to its non-rigidity and forces exerted by the subject. Unfortunately, while existing works [HCTB19, ZPJ+20, WY21, ZZB+21, GMSPM21] use human-environment contact and inter-penetration constraints to avoid collisions, they simply disregard such scene deformations, which results in substantial 3D reconstruction errors.

We argue that 3D scene deformations cannot be ignored if we would like to raise the accuracy of the reconstructed 3D human poses to the next level. Hence, this paper proposes *MoCapDeform*, a new framework for 3D monocular human motion capture with a 3D scene prior (a mesh); see Figs. 5.1 and 5.2 for an overview. In contrast to previous works, our method accurately localises the global human position in the scene from RGB inputs. It does so by a raycast contact finding policy to tackle the scale ambiguity problem, and a scene deformation modelling technique to address the limitation of low 3D reconstruction accuracy caused by the negligence of scene deformations.

Our method comprises three stages. In the **first stage**, we initialise the 3D human poses parameterised by the SMPL-X model [PCG+19], which can be done by any off-the-shelf 3D human pose estimator. This yields initial root-relative 3D poses that are reasonably accurate and sufficiently satisfy the image observations. Next, contact probability maps are estimated from the initial 3D poses, indicating which vertices on the human mesh are in contact with the 3D environment. The contact points are then re-projected to the image domain and passed through a raycasting operation to find the corresponding contact locations on the scene mesh. In the **second stage**, we register the estimated contact points on the human mesh to the raycasting results, leading to coarse global 3D human poses. **Finally**, these poses are further refined by jointly optimising for pose updates and deformations of the scene with which the body is in

contact.

In summary, our **contributions** are as follows:

- MoCapDeform—the first framework for joint markerless 3D human motion capture from monocular RGB images and capture of non-rigid 3D scene deformations. Such joint reasoning increases the accuracy of 3D human pose estimation on various benchmarks compared to existing methods (Sec. 5.4).
- A new raycasting based optimisation algorithm for finding dense contacts between humans and the environment (Sec. 5.3.3).
- A joint scene deformation and human pose refinement optimisation to recover both accurate human poses and scene deformations (Sec. 5.3.4).
- A new dataset for the experimental evaluation with human-scene interactions and observable scene deformations (Sec. 5.4.1).

We compare MoCapDeform to several previous state-of-the-art methods that assume monocular RGB images or videos and pre-scanned scene meshes input [HGT+21b, HCTB19, PCG+19]. Our approach regresses significantly more accurate global 3D human poses on the PROX [HCTB19] and our new datasets, producing reasonable scene deformations (Sec. 5.4). The new dataset and source code are available at `https://github.com/Malefikus/MoCapDeform`.

5.2 Related Works

Kinematic Monocular 3D Human Pose Estimation. Most works on monocular 3D human pose estimation reconstruct human joint positions in local coordinates. Some methods first estimate 2D poses in the 2D image space and then lift them into 3D [CR17, MHRL17, TRA17, MN17, FXW+18, DGM+19]. Several other approaches learn feature representations for 2D poses (without explicit 2D joint outputs) and perform lifting [MSS+17, PZDD17, HXM+19]. Further approaches regress 3D joints directly from RGB images via neural networks [TKS+16, MRC+17, RSF18]. While straightforward, this direct joint position representation has some issues, such as being hard to use in graphics applications, temporal jitter and implausible human skeletons due to varying bone lengths. These limitations can be addressed by estimating parameters of pre-defined body models such as joint angles for kinematic skeletons [ZSZ+16, MSS+17, MSM+20], pose and shape parameters of parametric body models [BKL+16, KBJM18, PZZD18, PCG+19, KAB20, ZHH+21], or template-based human performance capture methods [HXZ+19, XXG+20, HXZ+20, HXZ+21]. While most works estimate root-relative 3D human poses, several ones attempt to regress 3D human poses with absolute depths in the camera space [MCL19, SAA+20, MSM+20]. Without depth priors, however, it is difficult to obtain accurate and artefact-free human poses in the global reference frame.

3D Human Pose Estimation with Scene Constraints. While significant progress in 3D human pose estimation was made over the last decades [Gav99, MHK06, Pop07, SBIK16], utilising scene constraints remains insufficiently explored [HCTB19, RGH+20, ZZB+21, RBH+21, DSJ+21, SGL+22]. Several works consider the ground plane only, and by enforcing volume occupancy exclusions or detecting foot-floor contacts, they impose physical plausibility on the reconstructed motions [ZMS18, SAA+20, RGH+20, SGXT20, RBH+21]. Some works consider holistic representations and model complex scenes by placing arranged object meshes (recovered from categorised templates) into the desired coordinate frame [MGC+19, ZPJ+20, WY21]. This way, the knowledge about the spatial arrangement can be utilised to coarsely constrain the

global position of the human. Another line of work employs pre-scanned meshes of the whole environment [HCTB19, ZZB+21, WCR+22, CGM+20]. Hassan *et al.* [HCTB19] attempt to register empirically assumed contact vertices on the human body to the nearest scene vertices for human-scene collision detection. This approach does not fully resolve the depth ambiguity since the exact contact locations in the scene are still unknown. Zhang *et al.* [ZZB+21] use RGB-D videos as inputs and focus on the motion plausibility on top of scene constraints. All in all, the assumption of an available mesh enables accurately locating the human in the global reference frame, and provides the opportunity to model scene deformations for more accurate scene reconstruction and global human pose recovery. Considering the form of inputs (monocular RGB images + pre-scanned scene meshes) and outputs (global human poses), the PROX approach [HCTB19] is most related to our method. Another recent paper proposes HULC, *i.e.,* a framework for 3D human motion capture in complex environments [SGL+22]. They estimate dense human-scene contacts with a neural network trained on a new large-scale dataset, and not a learning-free raytracing like we do. In contrast to previous and concurrent works [HCTB19, ZZB+21, SGL+22], MoCapDeform models scene deformations, which helps to accurately locate humans in the global 3D space.

5.3 Method

We describe our optimisation framework with three stages; see Fig. 5.2. The inputs to our framework are RGB images of a static camera and a pre-scanned mesh of the scene at the beginning of capture aligned to the camera coordinate frame; the outputs are posed 3D human bodies and deformed scene meshes. The **first stage** (Sec. 5.3.2) performs *initial pose estimation*, where we estimate the initial pose from an RGB image, which suffers from scale ambiguity but faithfully overlays onto the images, *i.e.,* the root-relative pose is coarsely accurate, on top of which the contact probability map can be estimated. The **second stage** (Sec. 5.3.3) is a *global pose optimisation*, in which we utilise the estimated contact points on the human mesh, then cast camera rays through the human contact points to the scene mesh to find the contact points on the environment, and then optimise for the global poses respecting these contacts. Finally, in **stage three** (Sec. 5.3.4), we perform *joint scene deformation and pose refinement* to obtain accurate global 3D human poses and realistic scene deformations.

5.3.1 Assumptions and Notations

Our method assumes that a 3D mesh of the scene and its registration in camera space are given. The 3D mesh is represented by $M_s = (\mathbf{V}_s, \mathbf{F}_s)$, with vertices $\mathbf{V}_s \in \mathbb{R}^{N_v \times 3}$ (N_v as the number of vertices) and triangular faces $\mathbf{F}_s \in \mathbb{N}^{N_f \times 3}$ (N_f as the number of faces) containing the vertex indices of each triangle. The static 3D scene mesh can be reconstructed with standard commercial solutions, either leveraging Structure Sensor [str] and the Skanect [ska] software [HCTB19], or multi-view reconstruction and differentiable rendering techniques from Agisoft Metashape [Met21]. The reconstructed meshes contain surface normals that correctly indicate the "outside" and the "inside" of the scene. Based on the topology of the pre-defined scene mesh, our method outputs a deformed per-frame scene mesh $M'_s = (\mathbf{V'}_s, \mathbf{F}_s)$ (we omit frame indices in this notation for conciseness).

Following the representation of [HCTB19], we adopt a parametric 3D body model SMPL-X [PCG+19]. The model is formulated as a differentiable function $M_b(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\gamma})$ parameterised by shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$, body, hand and jaw pose $\boldsymbol{\theta} \in \mathbb{R}^{52 \times 3}$ of axis-angle representation with three

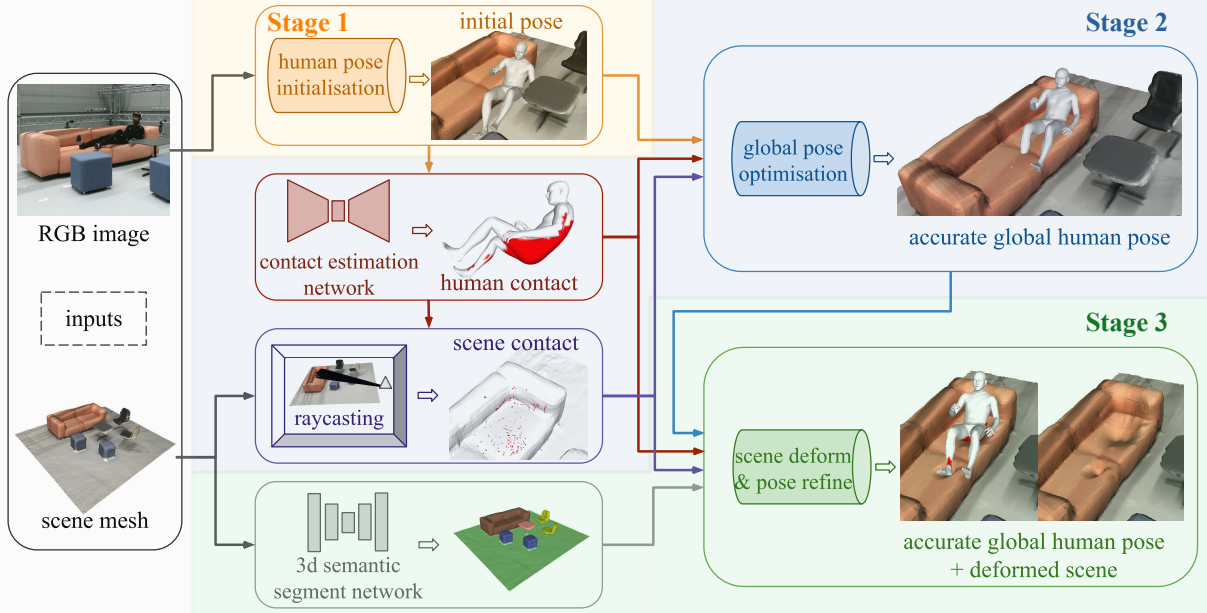


Figure 5.2: **Overview of MoCapDeform.** We first initialise the human pose and use it to find the contact points on the human mesh. Then, we apply raycasting to find the contact points on the scene mesh surface, which are then used to recover improved global human poses. Finally, we perform joint scene deformation and human pose refinement and obtain accurate global human pose and realistic scene deformations.

degrees of freedom (DoF) for each joint, holistic facial expression parameters $\boldsymbol{\psi} \in \mathbb{R}^{10}$, and global translation $\boldsymbol{\gamma} \in \mathbb{R}^3$ defining the global position of the pelvis joint. The model output is a 3D human mesh $M_b = (\mathbf{V_b}, \mathbf{F_b})$, with vertices $\mathbf{V_b} \in \mathbb{R}^{10475\times3}$ and the corresponding triangular faces $\mathbf{F_b} \in \mathbb{N}^{20908\times3}$ expressing the mesh connectivity. From the mesh vertices, we can regress an underlying rigged skeleton $J(\boldsymbol{\beta})$ with 55 joints defined by linear blend skinning. Following the notation of [BKL+16], we denote the posed joints as $R_{\boldsymbol{\theta\gamma}}(J(\boldsymbol{\beta})_i)$ for each joint i, where $R_{\boldsymbol{\theta\gamma}}$ denotes the kinematic tree defined by the pose parameter $\boldsymbol{\theta}$ and translation vector $\boldsymbol{\gamma}$. With this representation, we obtain a globally posed and shaped human body, *i.e.,* both the skinned mesh and the underlying articulation.

5.3.2 Stage1: Initial Human Pose Estimation

Stage 1 initialises the global human poses from monocular RGB images. This can be done by any of the off-the-shelf 3D human pose estimators [BKL+16, MSS+17, KBJM18, MSM+20, HCTB19, KAB20], which take the RGB images as input, follow closely the image cue and produce the required root-relative poses. As MoCapDeform is not restricted to sequential inputs, we employ a single-frame method for the initialisation. For a fair comparison with the SOTA approaches [HCTB19, HGT+21b], which are both based on the optimisation-based SMPLify-X [PCG+19], we initialise with SMPLify-X. Stage 1 minimises the following objective function:

$$E_1(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = E_J + \lambda_\theta E_\theta + \lambda_\alpha E_\alpha + \lambda_\beta E_\beta. \tag{5.1}$$

E_J is the RGB data term, *i.e.,* is a re-projection loss seeking to minimise the robust weighted distance between 2D joints—estimated from the RGB image using OpenPose [WRKS16, CSWS17, SJMS17]—and the 2D projection of the corresponding posed 3D joints of SMPL-X.

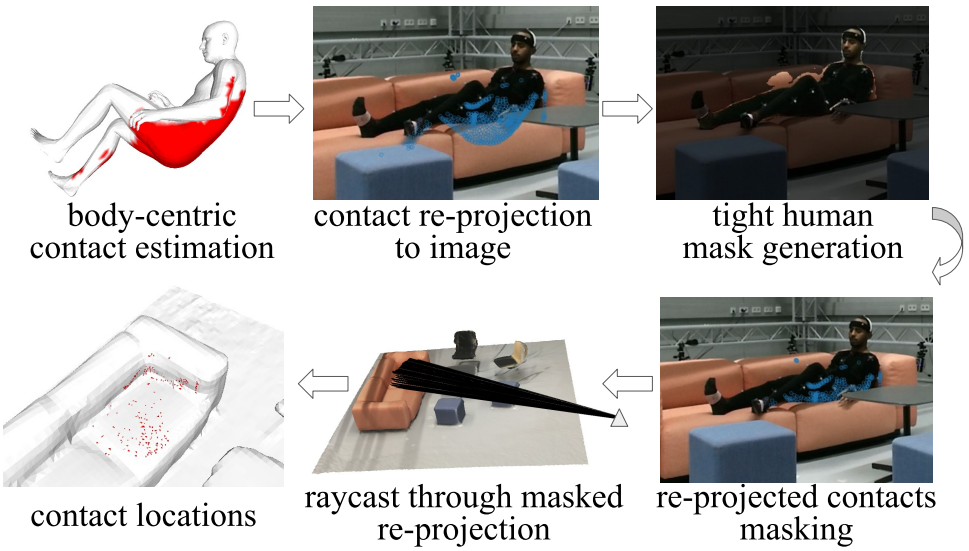


Figure 5.3: Overview of our raycast contact policy.

We assume a perspective camera model. The re-projections are weighted by the detection confidence scores, and a robust Geman-McClure error function [GEM87] is applied on top to down-weight noisy detections. E_θ is the trained VAE-based body pose prior of VPoser [PCG$^+$19], which enforces natural human poses learned from a large 3D human pose corpus [MGT$^+$19]. $E_\alpha = \sum_{i\in(\text{elbows,knees})} \exp(\boldsymbol{\theta_i})$ is a prior penalising extreme bending for elbows and knees. E_β is an ℓ_2-regulariser for human shape to penalise deviation from the neutral state. For a more detailed explanation, please refer to [PCG$^+$19, HCTB19].

5.3.3 Stage 2: Global Pose Optimisation

The goal of Stage 2 is to regress an accurate global 3D position of the human body estimated in Stage 1. Since monocular 3D human pose estimation is ill-posed without further priors, it is unlikely to obtain accurate results. Hence, we use the given scene mesh as a constraint to tackle depth ambiguity. We utilise the human-scene contact information by firstly estimating human body contacts on the initialised human body. We then perform raycasting of the re-projected human body contacts into the 3D scene to find the corresponding scene contacts and, finally, register the human body contact points to the scene contacts.

5.3.3.1 *Raycast Contact Estimation*

Our raycasting policy has three steps illustrated in Fig. 5.3: 1) Body-centric contact estimation, 2) Contact re-projection with masking, and 3) Raycast and scene contact estimation.

To find contacts on a scene through our raycast policy, we first need to find contacts on the human body. We thus employ POSA [HGT$^+$21b], *i.e.,* a conditional variational autoencoder (cVAE) that generates probability maps for different canonical human poses. The learned cVAE decoder takes as input human mesh points (in a root-relative pose and a canonical reference frame) and a latent vector $\boldsymbol{z}\sim\mathcal{N}(0, I)$ that conditions the sampling and can be directly applied to the initial human poses from Stage 1 since the root-relative poses are accurate. Specifically, we canonicalise the initialised human meshes from Stage 1 following the same formulation as in POSA and choose a zero vector $\boldsymbol{z}$ to generate contact probability samples that lie in the

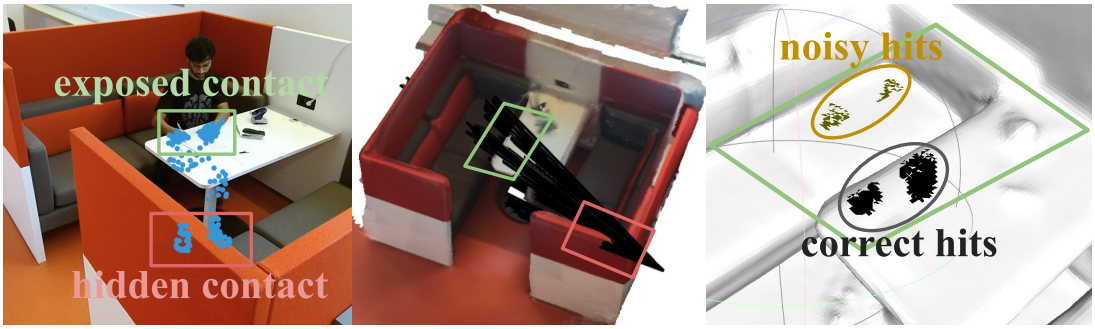


Figure 5.4: Detection of non-occluded areas and noisy contact label filtering based on the analysis of the ray-mesh hits.

peak in the learned latent space. Then, by thresholding the generated probability map by 0.5 [HGT+21b], we obtain the human mesh vertices that are in contact with the environment.

The generated human contacts are then re-projected to the image space. Then, we cast rays from the camera through these re-projections to find intersections with the 3D scene mesh. However, as the scene geometry is complex, there are usually multiple hits along each ray, and the challenge is to find out which hit is at the contact. Intuitively, as illustrated in Fig. 5.4, when the re-projected contacts fall on the body parts that are not occluded in the image (the green rectangles in Fig. 5.4), the front-most hits will be at the correct contacts. Otherwise, the rays firstly hit occluders (red rectangles in Fig. 5.4).

We empirically set the scanning radius ϵ=0.5 and MinPts=50 of the DBSCAN, to help eliminate the clusters (hit by the ray) that are far away from the main cluster or contain a small number of samples. After these steps, the resulting points are considered the corresponding scene contacts.

5.3.3.2 *Our Objective Function (Stage 2)*

With the help of the raycast results, we perform Stage 2 optimisation and register the estimated human contacts to the raycasted scene contacts. This leads to refinement of the initial estimates from Stage 1. We minimise the following objective function:

$$E_2(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, M_s) = E_J + \lambda_{\text{obs}} E_{\text{obs}} + \lambda_{\text{un}} E_{\text{un}} + \lambda_{t\theta} E_{t\theta} + \lambda_{t\gamma} E_{t\gamma}, \tag{5.2}$$

where E_J is as is in (5.1). E_{obs} is the "seen" contact term, which minimises the distance between estimated contact points on the human mesh and the raycasted contact points on the scene mesh. E_{un} denotes the "unseen" contact term, which registers the rest of the estimated human contacts that do not have raycast hits to the corresponding nearest scene vertices by minimising their Chamfer distance. In E_{obs} and E_{un}, the distances are calculated by a Geman-McClure error function [GEM87] to downweight noisy detections. Finally, we apply temporal smoothness terms $E_{t\theta}$ and $E_{t\gamma}$.

5.3.4 Joint Scene Deformation and Pose Refinement

In Stage 3, the coarse global poses obtained from Stage 2 are further refined, taking into account scene deformations. This stage jointly optimises for plausible 3D scene deformation in interaction regions and more accurate human poses, also resulting in fewer inter-penetrations between the two.

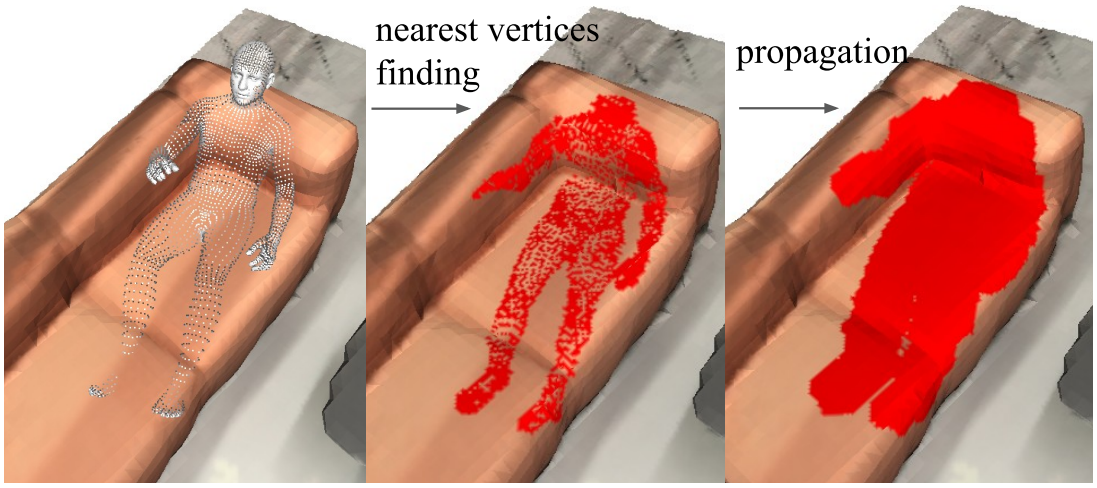


Figure 5.5: Determination of movable scene points.

5.3.4.1 *Scene Deformation Modelling*

To model deformations of non-rigid objects such as couches or beds, we use an as-rigid-as-possible (ARAP) regulariser [SA07]. It allows deforming a mesh with guidance by pre-defined sparse control vertices $c \in \Omega$ (Ω is a subset of contact point indices). The latter are first moved to the target positions, and the neighbouring faces are encouraged to stay rigid as much as possible. The deformed state of the entire surface is then found by optimising

$$E_{\text{ARAP}}(M'_s, \mathbf{R}_c) = \sum_{c\in\Omega} \sum_{n\in\mathcal{N}(c)} w_{cn} \left\| (\mathbf{v'}_c - \mathbf{v'}_n) - \mathbf{R}_c(\mathbf{v}_c - \mathbf{v}_n) \right\|, \tag{5.3}$$

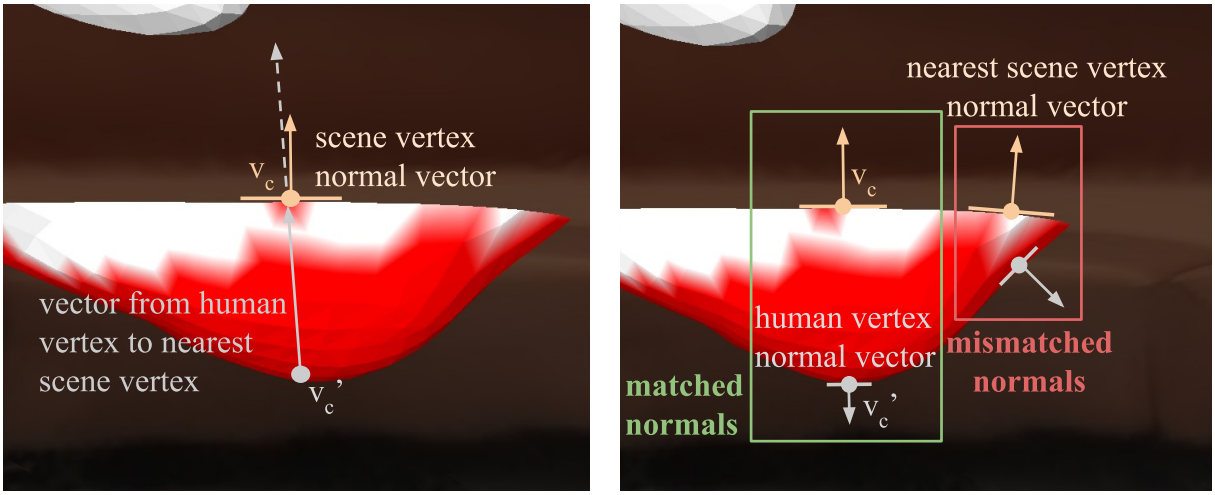
where $\mathbf{R}_c$ is the unknown rotation matrices; $\mathbf{v}_c$ and $\mathbf{v'}_c$ are the control vertex positions before and after the optimisation; $\mathcal{N}_c$ is the set of per-vertex neighbours $\mathbf{v_c}$; $\mathbf{v}_n$ and $\mathbf{v'}_n$ are the neighbouring vertices of $\mathbf{v}_c$ before and after optimisation, and w_{cn} are cotangent weights.

To integrate ARAP regulariser in our framework, we define on the scene mesh: 1) A set of control vertices $\mathbf{v_c}$; 2) The corresponding target positions $\mathbf{v'}_c$ for the control vertices; 3) A set of neighbouring vertices $\mathbf{v_n}$, which are allowed to move.

In practice, with the help of the current state of the human body, we first partition the whole scene mesh into movable and static areas and then choose the control points from the movable area and define the target positions accordingly. In the beginning, we need to know which parts of the scene are deformable and which are not. For that purpose, we adopt a 3D scene mesh segmentation network VMNet [HBS+21] trained on a large-scale indoor dataset ScanNet [DCS+17]. VMNet estimates semantic labels of the furniture. In this paper, we regard "sofa" and "bed" as non-rigid and all other object classes as rigid. After masking out rigid parts, we further define "movable" areas inside the deformable areas. As illustrated in Fig. 5.5, for every vertex on the human mesh, we find its nearest scene vertex and then propagate the obtained points to their k-th order neighbours. These points are regarded as "movable" during ARAP deformation. Here, we empirically set k=3 according to the scale of the meshes.

Next, we define the control vertices $\mathbf{v}_c$ on the scene mesh and their corresponding target positions $\mathbf{v'}_c$. Assuming that the human induces the scene deformations, $\mathbf{v}_c$ and $\mathbf{v'}_c$ can be defined according to the vertices $\mathbf{v}_{c'}$ of the human mesh (and their positions) that satisfy the following conditions:

- $\mathbf{v}_c$ is the nearest scene vertex to the human vertex $\mathbf{v}_{c'}$.



(a) collision checking (b) normal matching

Figure 5.6: Collision checking and normal matching. The viewpoint is "inside" the couch, looking at the colliding hip.

- $\mathbf{v}_{c'}$ are marked as "in contact" by the contact estimation step.
- $\mathbf{v}_{c'}$ collide with the scene mesh. Note that in this step, collision detection cannot be achieved by the pre-computed SDF since the scene is deforming. Hence, we check the collision status with the help of scene surface normals, as shown in Fig. 5.6a: When the vector from $\mathbf{v}_{c'}$ to $\mathbf{v}_c$ (grey) goes in the same direction as the normal vector of $\mathbf{v}_c$ (orange), $\mathbf{v}_{c'}$ will be classified as colliding with the scene.
- The normals of $\mathbf{v}_{c'}$ should be of opposite directions to the normals of $\mathbf{v}_c$, as is shown in Fig. 5.6b. This is because the scene usually deforms along the direction of the forces applied by the human.

At last, the nearest scene vertices $\mathbf{v}_c$ to the above human vertices $\mathbf{v}_{c'}$ are chosen as control points, and the control target positions $\mathbf{v'}_c$ are defined by the positions of $\mathbf{v}_{c'}$.

5.3.4.2 *Optimisation (Stage 3)*

The final stage is a joint and alternating optimisation for scene deformation and human poses. In every iteration k, we firstly pick the control points of ARAP according to the current human pose using the techniques described in Sec. 5.3.4.1 and then update the scene mesh using (5.3). Next, we update the human pose based on the updated scene mesh by minimising the following energy function:

$$E_3(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, M_s^k) = E_2(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, M_s^k) + \lambda_{\text{pen}} E_{\text{pen}}. \tag{5.4}$$

where $E_2(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, M_s^k)$ is defined in (5.2) and E_{pen} is a penetration term that does not use the pre-computed SDF, since the scene mesh is being constantly updated. Instead, it utilises the normal checking technique presented in Fig. 5.6a to detect collisions and then registers the colliding vertices on the human mesh to their nearest scene mesh vertices by minimising the Geman-McClure error function [GEM87].

5.3.5 Implementation

In Stage 1, closely following [PCG+19, HCTB19], we optimise (5.1) using a PyTorch [PGC+17] implementation of the limited-memory BFGS optimiser [NW06] with line search satisfying strong Wolfe conditions. (5.2) in Stage 2 and (5.4) in Stage 3 are optimised by PyTorch implementations of the Adam optimiser [KB15]. For the scene deformations, *i.e.,* (5.3) in Stage 3, we adopt the ARAP implementation from Open3D [ZPK18]. The off-the-shelf components we adopt (*i.e.,* PointRend [KWHG20], VMNet [HBS+21] and POSA human contact estimation [HCTB19]) are easily deployable and sufficiently accurate for our task. The λ and w_{cn} weights in (5.1)-(5.4) are empirically found and fixed in all experiments; see the source code.

5.4 Experiments

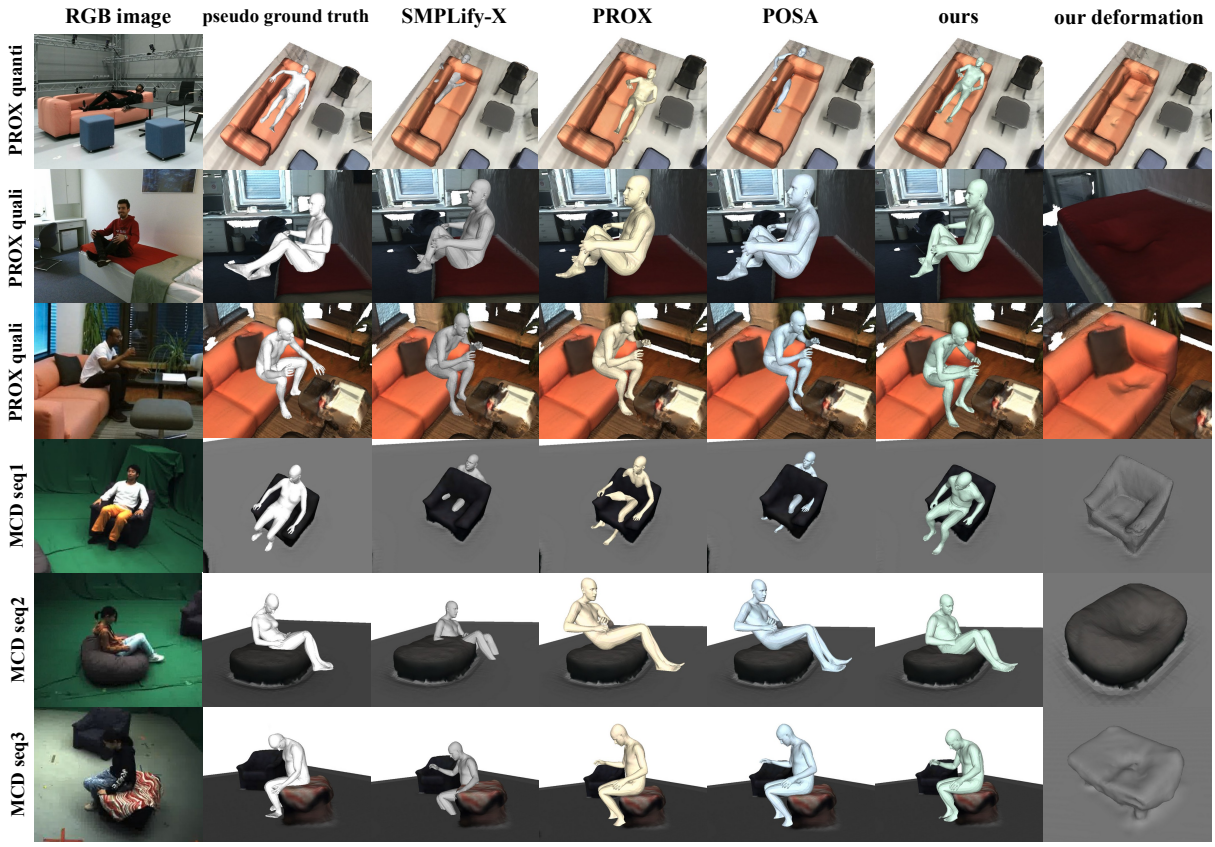


Figure 5.7: Qualitative results and comparisons on different datasets. Our MoCapDeform achieves more accurate global localisations than the state-of-the-arts, leads to less penetration, and prevents the human bodies from floating when there are large scene deformations. Moreover, it outputs plausible scene deformations not reconstructed by the previous methods.

To evaluate our MoCapDeform framework, we conduct extensive experiments on two datasets (Secs. 5.4.1-5.4.2) and show qualitative results (Sec. 5.4.3).

Table 5.1: Results on the PROX dataset using RGB inputs. We show our results of Stages 1 and 2 ("s1+s2") and full method and compare them with several state-of-the-arts. Best is indicated in **bold**, and second best in ***bold italic***.

	PJE	V2V	p.PJE	p.V2V	non-col
SMPLify-X [PCG+19]	214.64	219.20	64.04	61.88	92.42%
PROX [HCTB19]	167.16	171.35	63.54	63.06	95.63%
POSA [HGT+21b]	157.11	159.52	63.70	63.23	95.89%
MoCapDeform (s1+s2)	***144.15***	***145.23***	***62.86***	***61.19***	***95.90%***
MoCapDeform (full)	**139.78**	**140.60**	**62.29**	**60.67**	**97.60%**

Table 5.2: Results on MoCapDeform dataset using RGB inputs. We compare outputs of Stages 1 and 2 ("s1+s2") and our full method to several state-of-the-art approaches. Best is indicated in **bold**, and second best in ***bold italic***.

	PJE	V2V	p.PJE	p.V2V	non-col
SMPLify-X [PCG+19]	441.86	451.87	**89.73**	**101.53**	97.14%
PROX [HCTB19]	375.01	403.22	97.09	107.57	97.99%
POSA [HGT+21b]	365.91	398.15	97.26	108.67	98.41%
MoCapDeform (s1+s2)	***266.18***	***283.46***	***91.18***	***101.71***	***98.57%***
MoCapDeform (full)	**264.68**	**282.01**	91.91	102.43	**99.04%**

5.4.1 Datasets

PROX dataset [HCTB19]. The PROX dataset includes a large qualitative and a small quantitative sets. The qualitative set contains monocular videos of 20 human subjects interacting with 12 indoor scenes along with the 3D scene scans: altogether, $100k$ RGB-D frames recorded at 30 fps (without ground-truth 3D human poses). The quantitative set contains 180 static RGB-D frames, with one human subject wearing markers interacting with a mimicked living room containing daily furniture. The pseudo-ground-truth SMPL-X parameters for the quantitative set are fitted by the marker-based MoSh++ [MGT+19] method.

MoCapDeform (MCD) dataset. To evaluate all outputs of MoCapDeform, including the deformations, we record a new dataset of people interacting with furniture, *i.e.,* a non-rigid sofa, a deformable stool, and especially a beanbag, which retains its deformed shape after the interaction and allows obtaining ground-truth deformations. We reconstruct accurate human meshes and scene geometry with the multi-view camera setting and a markerless differentiable-rendering-based technique [Met21]. The human meshes can then be used to fit the SMPL-X model parameters and serve as ground truth. The dataset contains four video sequences at 30 fps of four subjects interacting with the furniture ($16k$ sequential RGB images in total). We utilise the dataset for both quantitative and qualitative experiments.

5.4.2 Quantitative Evaluation

We evaluate the estimated 3D human poses by computing several quantitative metrics indicating the global and local 3D reconstruction accuracy and the degree of penetrations. For global 3D reconstruction accuracy, we adopt the standard **PJE** and **V2V** metrics and report them in mm.

Table 5.3: Results on the PROX quantitative dataset using RGB-D inputs. Best is indicated in **bold**.

	PJE	V2V	p.PJE	p.V2V	non-col
SMPLify-D [HCTB19]	70.63	72.19	44.58	44.33	93.65%
PROX-D [HCTB19]	63.03	65.64	39.89	39.74	93.86%
POSA-D [HGT$^+$21b]	62.44	66.16	39.73	40.11	93.97%
MoCapDeform (s1+s3)	**59.32**	**62.37**	**39.57**	**39.12**	**97.04%**

PJE stands for the mean per joint position error, and V2V indicates the mean vertex-to-vertex error. For local 3D reconstruction accuracy, we employ **p.PJE** and **p.V2V** (in *mm*), which are the PJE and V2V metrics after Procrustes alignment. Furthermore, to evaluate the human-scene penetrations—following the work in this domain [HCTB19, ZZM$^+$20, ZHN$^+$20, ZZB$^+$21]—we report the **non-collision score**, which is the percentage of human body mesh vertices that do not penetrate the scene mesh. Note that for MoCapDeform, the non-collision score is calculated over the deformed meshes since they are also an output of our method.

The results on the PROX dataset and our new dataset are summarised in Tables 5.1 and 5.2. We report the results of Stages 1+2 and all stages of our framework (full model). For our new MoCapDeform dataset, we down-sample the framerate to 5 fps. As can be observed in the tables, both the global pose optimisation and the scene deformation stages contribute to more accurate pose estimation and outperform the previous approaches. Our method achieves significant improvement in terms of the global poses. With the help of the estimated scene deformations, the final output meshes of MoCapDeform have significantly fewer penetrations.

To further evaluate the effectiveness of the scene deformation stage, we conduct experiments on top of the RGB-D inputs from the PROX quantitative dataset; see Table 5.3. Specifically, we replace the pose initialisation stage with the PROX-D method [HCTB19], in which a depth term is used as a constraint during optimisation, supposedly resolving the depth ambiguity. Then we skip the second stage and directly apply our joint scene deformation and pose refinement stage over the PROX-D initialisation, with a depth data term (the same one as used in [HCTB19]) added to (5.4). The numbers in Table 5.3 show that our scene deformation stage can further improve the accuracy of 3D human pose estimation in all metrics and results in significantly fewer inter-penetrations after the deformation (*cf.* non-collision score).

5.4.3 Qualitative Results

We show qualitative results on PROX and our MoCapDeform datasets in Fig. 5.7. SMPLify-X [PCG$^+$19] inaccurately localises the human and causes severe penetrations, as it ignores scene information. Both PROX [HCTB19] and POSA [HGT$^+$21b] leverage scene constraints by penalising the human-scene collisions with pre-computed SDF values of the static scenes. Since they do not model scene deformations, the collision penalising terms tend to lift the human above the scene surfaces even in the presence of large scene deformations. This leads to the subject floating (Fig. 5.7, PROX: rows 1, 2 and 5; POSA: rows 2 and 5), causing inaccurate global positions along the depth channel or severe body-scene penetrations (Fig. 5.7, PROX: rows 3 and 4; POSA: rows 1, 3 and 4), as the image cues and anti-collision terms cannot be satisfied simultaneously. In contrast, with the help of our raycast contact algorithm and scene deformation modelling, MoCapDeform finds more accurate global human positions without the floating issue, with significantly fewer inter-penetrations, and outputs plausible scene deformations. See Fig. 5.8

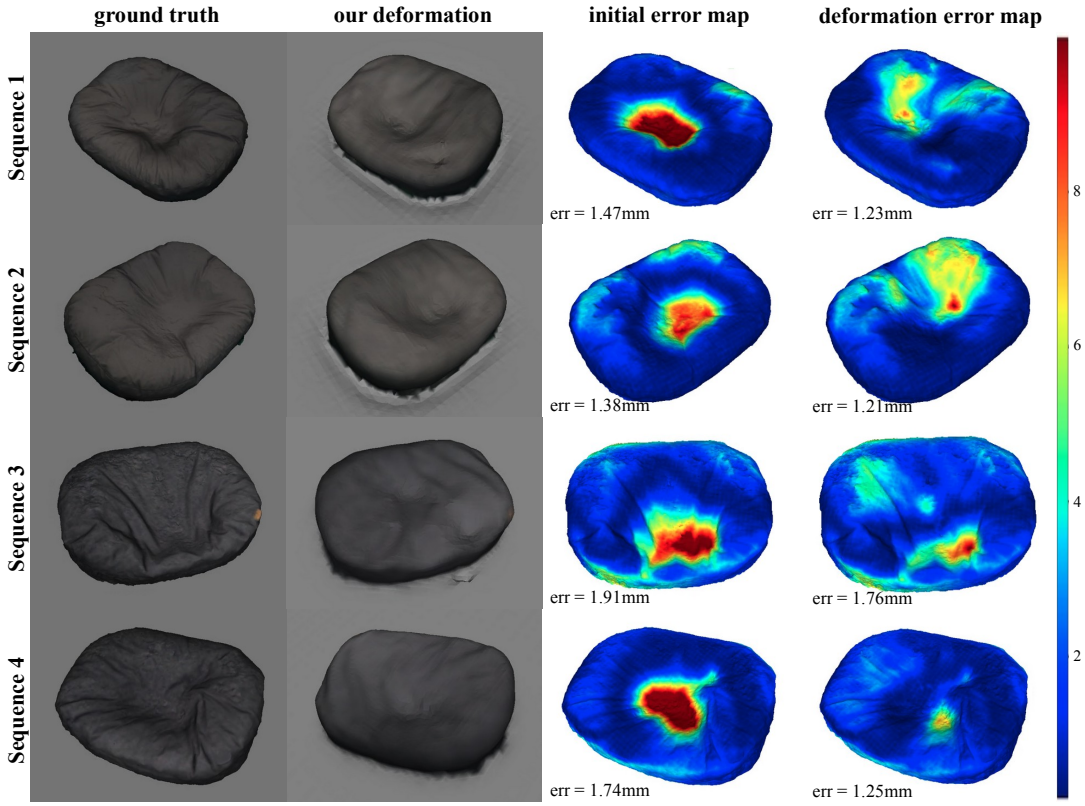


Figure 5.8: Comparison of ground-truth meshes and our deformations. The error maps show colour-coded per-vertex distances between the ground-truth meshes and the initial shapes or final states estimated by MoCapDeform.

for comparisons between the ground-truth states of the beanbag (reconstructed after the person stands up) and the deformation output from our method.

5.5 Conclusion

One limitation of our method is the dependency on reasonable 3D pose initialisation; recovery from starkly erroneous initial poses is unlikely. Moreover, severe scene occlusions blocking all human-scene contacts violate the assumptions of the raycast module, hindering the global pose optimisation. One future direction could be accounting for elastic properties of different objects and the integration of more fine-grained deformation models enabled by segmenting a single object into rigid and non-rigid parts. Next, modelling interactions between subjects wearing loose clothes and a non-rigid environment is an intriguing direction.

Closing Remarks. We present *MoCapDeform*, the first framework for markerless global 3D human motion capture from monocular RGB images with the awareness of non-rigid scene deformations. Benefiting from our new raycast-based contact localisation and joint scene deformation and pose optimisation steps, we find accurate global human poses and, at the same time, reasonable scene deformations. We show significantly improved global 3D human poses compared to several competing approaches. Due to these encouraging results, we expect in future to see more research on human motion capture systems that are aware of scene changes, including non-rigid deformations.

II

Reconstructing and Understanding 3D Scenes

6 Test-time Adaptation for Monocular Depth Estimation

Contents

Test-time domain adaptation, i.e. adapting source-pretrained models to the test data on-the-fly in a source-free, unsupervised manner, is a highly practical yet very challenging task. Due to the domain gap between source and target data, inference quality on the target domain can drop drastically especially in terms of absolute scale of depth. In addition, unsupervised adaptation can degrade the model performance due to inaccurate pseudo labels. Furthermore, the model can suffer from catastrophic forgetting when errors are accumulated over time. We propose a test-time domain adaptation framework for monocular depth estimation which achieves both stability and adaptation performance by benefiting from both self-training of the supervised branch and pseudo labels from self-supervised branch, and is able to tackle the above problems: our scale alignment scheme aligns the input features between source and target data, correcting the absolute scale inference on the target domain; with pseudo label consistency check, we select confident pixels thus improve pseudo label quality; regularisation and self-training schemes are applied to help avoid catastrophic forgetting. Without requirement of further supervisions on the target domain, our method adapts the source-trained models to the test data with significant improvements over the direct inference results, providing scale-aware depth map outputs that outperform the state-of-the-arts. Code is available at https://github.com/Malefikus/ada-depth.

This chapter is based on [LSSD23]. As the first author, Zhi Li worked on all parts of the project, including idea generation, all implementation and experiments, and paper writing.

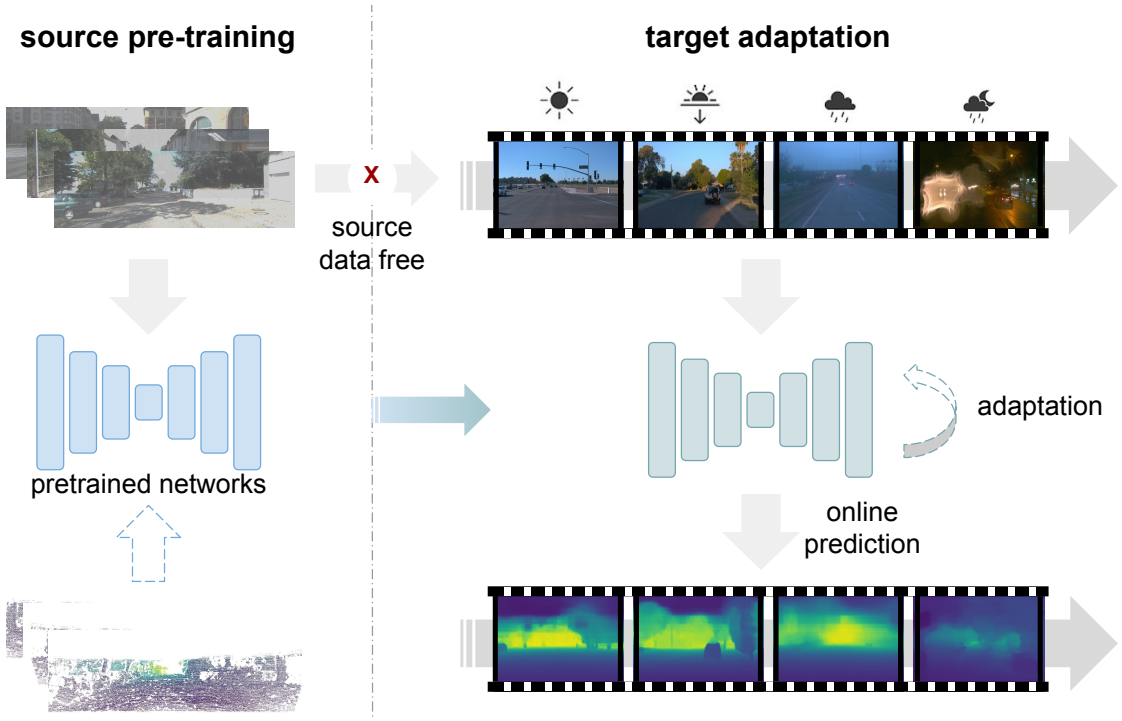


Figure 6.1: Overview of our test-time domain adaptation framework. We adapt our source-trained network to the changing target data during test time in an online fashion, without requiring the access of the source data anymore.

6.1 Introduction

Continuous depth estimation from monocular videos is one of the many fundamental tasks for the development of navigation systems, and is particularly important in autonomous driving scenarios. The development of deep learning enables the acquisition of neural networks learned from large training corpus which can then be used to make inference on the test data. In real world applications, however, the training data (source) are not always sufficient for the model to perform well enough on the test data (target), especially when the domain gap is large, e.g. when the camera setups change, when locations and weather conditions are changing, etc. To tackle this issue, domain adaptation techniques are introduced [KPVG21, WFVGD22].

Specifically, to acquire better inference from the source-trained model, one needs to further fine-tune the network on the test data in an unsupervised or self-supervised manner, and more practically in an online fashion, i.e. making prediction on the current frame directly after the adaptation to this frame, then moving on to the next frame. In addition, due to privacy concerns, legal constraints or technical reasons, the source data are generally considered unavailable during the adaptation process, making it more challenging but expanding its applicability. In general, source-free test-time domain adaptation is crucial and practical to real world machine perception applications.

We tackle the problem of domain adaptation for depth estimation under this challenging source-free, online setup with continuously changing environment. Fig. 6.1 is a brief recap of our framework. On the source domain, we train networks on the training videos and the corresponding Lidar groundtruth depth maps. To adapt the source-trained networks to the target domain, we fine-tune the source-pretrained networks with the aligned test images consecutively in

a temporal order on-the-fly. Our framework is designed for stable, long-term adaptation without catastrophic forgetting but still with a lot of accuracy improvement.

In summary, our contributions are as follows:

- We develop a test-time domain adaptation framework for continuous depth estimation from monocular videos that significantly improves the inference quality of source-trained models on target data and outperforms the state-of-the-arts. Unlike previous works, our framework is able to produce scale-aware depth predictions on the target data without requiring additional supervision from either target domain or source domain, thanks to our novel 3-branch network taking advantage of both supervised and self-supervised models.
- We propose a simple yet effective pixel scale alignment scheme between source and target data based on geometrical constraints, which significantly improves the source-model inference quality on the target domain already before the adaptation.
- A novel consistency checking technique is proposed to filter out erroneous pixels in the pseudo labels during adaptation, which improves the pseudo label quality thus enhances the adaptation performance.
- We enable long-term adaptation without catastrophic forgetting by proposing an effective regularisation scheme integrated into the effective EMA self-training scheme in the adaptation process.

6.2 Related Works

6.2.1 Domain Adaptation

Domain adaptation, or more specifically referred to as unsupervised domain adaptation (UDA), aims to improve the model trained on (usually labelled) source domain for the inference on the unlabelled target domain [PTKY10, PGLC15]. There are various techniques for adaptation, such as input alignment [HTP+18, YS20] and feature distribution alignment between two domains [LCWJ15, GL15, THS+18]. In addition to feature alignment, there are self-training techniques [HDVG22, LLDG19, WDH+21, ZYKW18] which fine-tune the model to the target domain using pseudo labels created by the model itself. In general, domain adaptation is still task-specific – feature alignment and self-training become challenging for dense predictions such as depth estimation, and there are not a lot work which aims at this specific task [KPVG21, TPMDS19]. Under different additional constraints, there are different variants of domain adaptation tasks.

Test-Time Domain Adaptation test-time domain adaptation adds a constraint to the general domain adaptation by restricting the access to the source domain data during adaptation [KVB+20, YWvdW+21]. Both feature alignment and model fine-tuning becomes challenging in the absence of source data. Generative models are utilised in some works to support feature alignment on target data [KSN21, LJC+20, YYYH21]. Other works fine-tune the source model to the target domain without explicitly performing feature alignment, but by updating the trainable parameters in BatchNorm layers using entropy minimisation [WSL+20, LHF20, KECK21, ZL21, HUC+21, YLZ21]. While showing promising results, standard test-time domain adaptation still assumes an offline scenario where all the test data are provided for network fine-tuning, making the applicability limited.

Continuous Domain Adaptation continuous domain adaptation does not limit the target domain to a specific one, but assumes continually changing target data [WFVGD22]. Existing

works in this line achieve continuous adaptation by techniques such as adversarial feature alignment [WBP18] or [BTHD18]. Most of the works, however, require access to data from both the source and target domains in order to align the distributions. [WFVGD22] tackles the problem of source-free, continuous test-time domain adaptation, but only for classification task. For the specific task of depth estimation under this setup, there is not a lot research done, except for [KPVG21] which still requires additional supervision (ground truth velocity) from the target data in order to generate correctly scaled depth maps.

6.2.2 Monocular Depth Estimation

Monocular depth estimation aims to predict dense depth maps from single-view RGB images or videos. A lot of research is done in training neural networks by either supervised or self-supervised means.

Supervised Monocular Depth Estimation supervised monocular depth estimation methods assume the availability of paired input images and groundtruth Lidar points and utilise the sparse Lidar points as supervisory signals to train models that can produce dense depth maps. One of the earliest work in this line is [EPF14], in which an end-to-end encoder-decoder network is developed to learn dense depth maps from sparse Lidar groundtruth. The line of research is continued with improvements on network architectures [EPF14, LHKS19, AVIL21, LLK$^+$21, SCC$^+$22, YGD$^+$22, LWLJ22]. and the use of temporal cues [ZSL$^+$19] or cross-modal consistencies such as semantics and normals [ZCX$^+$19, YLSY19]. However, the high cost of Lidar makes it hard to assume the groundtruth to be always available, which is a significant drawback of the supervised methods.

Self-supervised Monocular Depth Estimation Self-supervised monocular depth estimation methods generally formulate the depth estimation task as a view synthesis problem [ZBSL17, GMAFB19], where depth and relative camera pose estimators are trained jointly to calculate the view synthesis losses. Improvements are done on top of this basic training framework, for example the utilisation of geometrical consistency through auxiliary optical flow estimations [CSS19, LHS$^+$20], additional weak supervisions such as velocity [GAP$^+$20], or temporal cost volume calculation [WMAP$^+$21]. Although self-supervised methods achieve promising results without requirement of groundtruth Lidar points during training thus can be easily deployed in domain adaptation scenarios, it has a fatal drawback – the output depth maps are always up to an unknown scale, meaning that we still need the groundtruth Lidar points of the test data to apply scale alignment during inference, unless we develop reliable scale recovery techniques (such as in [XZH$^+$20]), which is non-trivial.

6.3 Method

In this section we show the detailed design of our test-time domain adaptation framework for continuous depth estimation. Given monocular video data with corresponding groundtruth Lidar points in the source domain, we first train two source models using existing supervised and self-supervised depth estimation training scheme, respectively. Note that the supervised model is able to produce absolute scale depth maps, while the self-supervised model does not have any scale awareness, by principle. During adaptation, as is shown in Fig. 6.2, we first initialise three branches (a regularisation branch, a supervised branch and a self-supervised branch) using the source-trained supervised model (for the first two) and the self-supervised model (for the last). When new data come, the self-supervised branch can be updated in the way it is originally trained, then be used to generate a pseudo label; the regularisation branch produces another

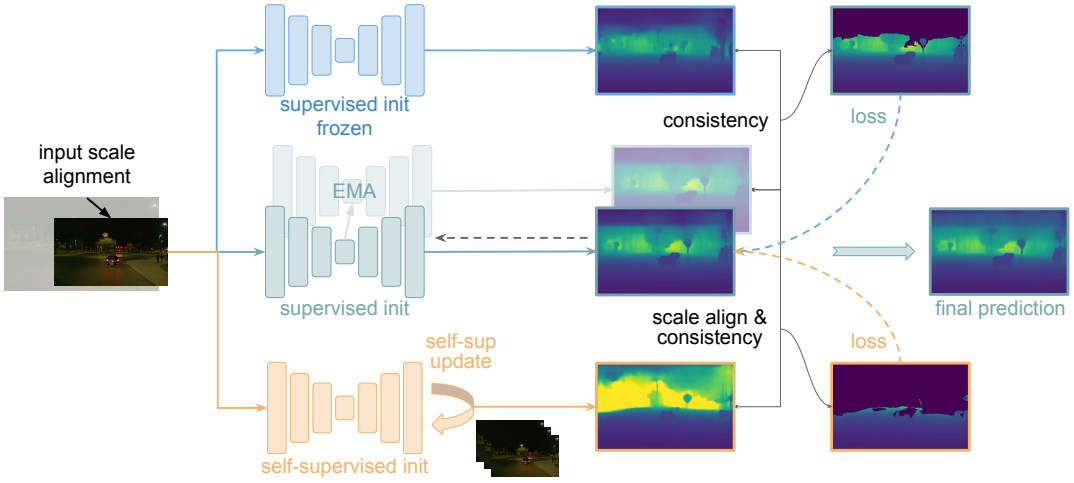


Figure 6.2: Pipeline of our adaptation framework. The three branches (from top to bottom) are initialised by source-trained supervised model/ supervised model/ self-supervised model, respectively. For every frame of the test data, the self-supervised branch (bottom) is firstly updated by the unsupervised image synthesis loss which requires only 2 adjacent RGB frames, then be used to create a pseudo label. The regularisation branch (top) generates another pseudo label. The supervised branch (middle) makes a prediction which is then compared with the two pseudo labels, to filter out less confident pixels and create more robust pseudo labels. These pseudo labels are used to update the supervised branch. To increase stability we adopt the EMA [TV17] self-training scheme for supervised branch. After the iteration, the supervised branch makes an accurate, scale-aware final prediction, and the networks move on to the next frame. Some network details are omitted for simplicity and will be introduced in the texts.

pseudo label for regularisation. The two pseudo labels are then compared with the prediction made by the supervised branch, to filter out less confident pixels. The filtered pseudo labels are then used to calculate losses thus update the supervised branch. The updated supervised branch makes final predictions for this frame, and the framework moves on to the next frame.

Our 3-branch adaptation framework is designed in a way that we can benefit from the knowledge distilled from both the supervised and the self-supervised models. Research [BKK22] has shown that disentangling the supervised and self-supervised training conduces the complementary advantages of both loss functions, while training only one model mixing both the losses results in inheriting the limitations from both. For our test-time domain adaptation problem, specifically, the supervised branch provides scale-aware, robust depth predictions which can be further improved by self-training techniques during adaptation, but the improvement is limited; the self-supervised branch generates scale-ambiguous depth estimation, but can be quickly adapted to the test data via self-supervised loss with significant performance improvement. With the two sets of pseudo labels created by the two branches, we improve the accuracy of depth predictions on the test data while retaining the scale-awareness and robustness, even for long-term adaptation.

6.3.1 Supervised Branch

For the supervised branch in our framework, we pretrain a single-frame monocular depth estimation network in a supervised way on the source data, following the network architecture in [YGD+22]. It consists of a powerful swin-transformer [LLC+21] encoder, and a hierarchical

decoder comprising four levels of neural window FC-CRF modules [YGD+22]. A supervised loss $\mathcal{L}_s$ is defined between the network predictions $\hat{d}$ and groundtruth sparse Lidar points d^*. Following common practices in previous works [BAW21, LHKS19, LLK+21, YGD+22], we adopt the Scale-Invariant Logarithmic (SILog) loss proposed by [EPF14]. First, the logarithm difference between $\hat{d}$ and d^* is calculated on each pixel i of the K pixels where the groundtruth Lidar points are available:

$$\Delta d_i = \log \hat{d}_i - \log d_i^*, \tag{6.1}$$

then the SILog loss is computed as:

$$\mathcal{L}_s = \alpha \sqrt{\frac{1}{K}\sum_i \Delta d_i^2 - \frac{\lambda}{K^2}(\sum_i \Delta d_i)^2}, \tag{6.2}$$

where λ is a variance minimising factor and α is a weight constant controlling the scale of the loss. Following previous works [LHKS19, YGD+22], we set λ to 0.85 and α to 10.

6.3.2 Self-Supervised Branch

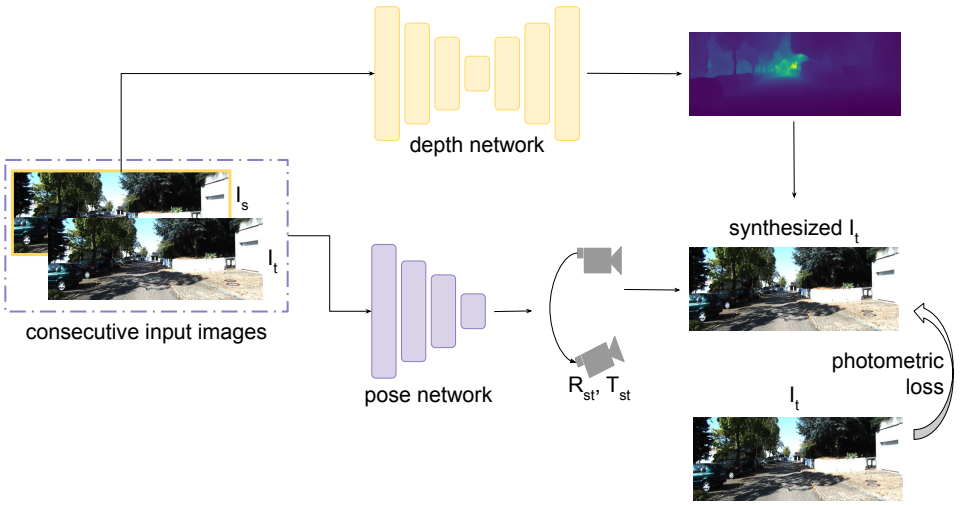


Figure 6.3: Self-supervised monocular depth estimation framework.

We train our self-supervised depth estimation branch on monocular source videos following the Monodepth2 [GMAFB19] pipeline, where two separate networks – a depth network and a pose network – are jointly trained. As is shown in Fig. 6.3, we jointly train two separate networks – a depth network and a pose network. For the network architectures we follow [GMAFB19] and adopt two separate ResNet50 [HZRS16] encoders and task specific convolutional decoders for each. The depth network receives an RGB image as input and produces its corresponding depth map; the pose network takes adjacent frames (in our experiments, 2 frames) as input and predicts the relative poses between the frames. The predicted depth map and relative poses can be used to warp one image frame to another; then, a loss function can be defined on the synthesised image and the target image by calculating their photometric differences. For more detailed explanations of the network architecture and losses please refer to [GMAFB19].

Specifically, assume that camera intrinsic matrix K is known in advance, given a source image I_s and its adjacent target image I_t captured by moving camera, let $M_{s\to t}$ be the camera

pose (extrinsics) that defines the 3D translation $T_{s\to t}$ and rotation $R_{s\to t}$ between consecutive 3D scene positions:

$$M_{s\to t} = \begin{bmatrix} R_{s\to t} & T_{s\to t} \\ 0 & 1 \end{bmatrix}. \tag{6.3}$$

Given a pixel p in the source frame, its corresponding position in 3D in homogeneous coordinates x can be computed by back projection using the predicted source depth:

$$x = \begin{bmatrix} D_s(p)K^{-1}p \\ 1 \end{bmatrix}. \tag{6.4}$$

x can then be re-projected to the target frame using the estimated camera pose $M_{s\to t}$, under the assumption of a static scene. The pixel p in the source frame is thus warped to the pixel p' in the target frame:

$$p' = [K|0]M_{s\to t}x. \tag{6.5}$$

The synthesised target image $I_{s\to t}$ can be obtained by sampling the source images I_s using bilinear interpolation [GMAB17, GMAFB19], denoted as $I_s\langle p'\rangle$.

Specifically, assume that camera intrinsic matrix K is known in advance, given a source image I_s and its corresponding depth map prediction, a synthesised target image $I_{s\to t}$ can be obtained by sampling the source image [GMAB17, GMAFB19]. Therefore, unsupervised loss $\mathcal{L}_u$ can be defined as a per-pixel photometric reprojection loss $\mathcal{L}_{pe} = pe(I_t, I_{s\to t})$ between the synthesised image $I_{s\to t}$ and the target image I_t which is adjacent to the source image, together with an edge-aware smoothness loss $\mathcal{L}_{sm}$ to enforce smoothness of depth maps, following the formulation in [GMAFB19]:

$$\mathcal{L}_u = \mathcal{L}_{pe} + \lambda\mathcal{L}_{sm}, \tag{6.6}$$

where λ is the weighting factor between the two terms (which is set to 10^{-3} for all the experiments, following [GMAFB19]).

6.3.3 Target Domain Scale Alignment

In domain adaptation scenarios, most of the times there are different camera setups between source and target data, resulting in different image sizes. Given that the real world scale does not vary, larger image sizes stand for larger object sizes on the image, which, by intuition, results in different scale estimations for the predicted depth map. We seek to align the pixel scales between the source and target images in order to fix the image/real world proportions, thus avoid drastic disparity between estimated scales on the source and target data. We assume that we do not have access to the source data, but only the metadata (camera intrinsics, camera height to the ground, etc.) of the source and target data.

According to principles of imaging [HZ03], larger focal length results in bigger object size on the image. Therefore, given a target image $I_t(H_t, W_t)$ with height H_t and width W_t whose unit is aligned with the focal length f_t, and a known source camera focal length f_s, we can align the target image "pixel sizes" (i.e. object sizes) to the source image by simply resizing the image by the proportion of the focal lengths:

$$\hat{I}_t = I_t(H_t \cdot \frac{f_s}{f_t},\ W_t \cdot \frac{f_s}{f_t}). \tag{6.7}$$

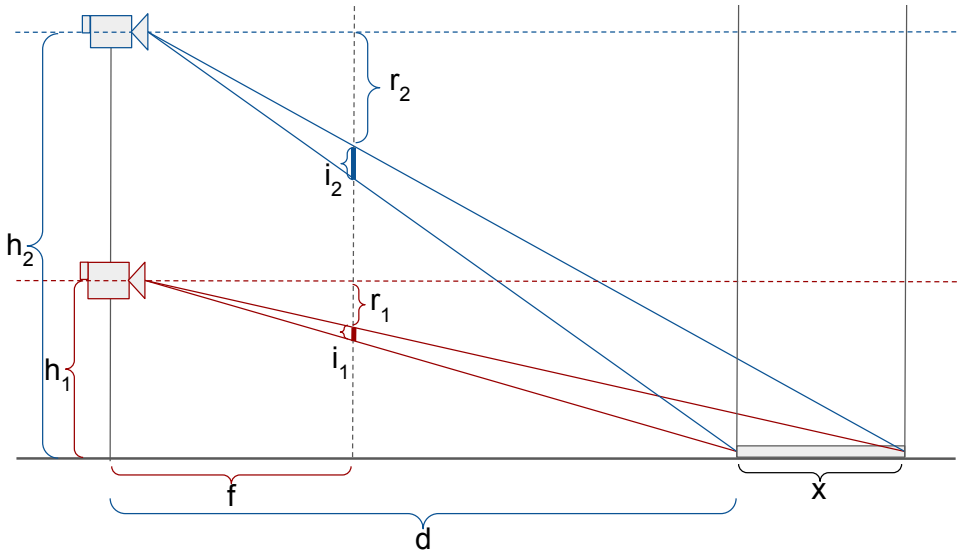


Figure 6.4: Imaging of the ground plane under different camera height.

6.3.4 Pixel Alignment with Camera Height

Another important factor which affects the scale of imaging is camera height. In autonomous driving scenes, the imaging of the stretch of the pavement is an important reference to absolute scale estimation; we seek to further align the pixel sizes of the roads on the images between source and target data, to avoid scale drifting during inference. Specifically, for a stretch of road with a certain distance to the camera, the two cameras whose heights are h_1 and h_2 share the same focal length (or the focal lengths are "aligned" for the images with Equation 6.7), the sizes of this stretch of road on the images are i_1 and i_2, respectively. According to principles of triangularity, we can derive that the sizes of the road on the images are proportional to the camera heights. We have:

$$\begin{aligned} \frac{i_1 + r_1}{f} &= \frac{h_1}{d}, \quad \frac{i_2 + r_2}{f} = \frac{h_2}{d}, \\ \frac{h_1}{d+x} &= \frac{r_1}{f}, \quad \frac{h_2}{d+x} = \frac{r_2}{f}, \end{aligned} \tag{6.8}$$

thus:

$$\frac{h_1}{h_2} = \frac{i_1 + r_1}{i_2 + r_2}, \quad \frac{h_1}{h_2} = \frac{r_1}{r_2}, \tag{6.9}$$

therefore the sizes of the road on the images are proportional to the camera heights:

$$\frac{h_1}{h_2} = \frac{i_1}{i_2}. \tag{6.10}$$

With Equation 6.10 we can further align the pixels of $\hat{I}_t$ to the source, given the camera height of the source data h_s and target data h_t.

The complete scale alignment process will be simply resizing the target images according to the focal lengths and camera heights:

$$\hat{I}'_t = I_t(H_t \cdot \frac{f_s}{f_t} \cdot \frac{h_s}{h_t},\ W_t \cdot \frac{f_s}{f_t} \cdot \frac{h_s}{h_t}). \tag{6.11}$$

6.3.5 Continuous Test-Time Adaptation

After the source models being trained and the target images sizes aligned, we perform our adaptation on the test videos using the pipeline described in Fig. 6.2. The supervised branch and self-supervised branch are initialised by the supervised and self-supervised models trained on the source data, respectively, and will be updated during adaptation; the regularisation branch is a copy of the source trained supervised model and is kept unchanged during the adaptation process. In every iteration, we first update the self-supervised branch with the self-supervised loss computed on this frame, then make prediction on this frame to be used as a pseudo label; the regularisation branch produces another pseudo label. Both pseudo labels are compared with the prediction of the supervised branch, to filter out less reliable pixels via consistency check. The filtered pseudo labels are then used to update the supervised branch. In addition, an EMA self-training scheme is adopted while updating the supervised branch, to further improve performance gain. The updated supervised branch makes final prediction for this frame, and the adaptation moves on to the next frame. Details of each component are listed below:

6.3.5.1 *Pseudo Label Consistency*

A consistency checking scheme is performed on the pseudo labels created by the self-supervised branch and the regularisation branch, which is conducted by computing the per-pixel difference between one pseudo label D_p (either from the regularisation branch or the self-supervised branch) and the corresponding prediction from the supervised branch D_s and masking out the pixels that have bigger difference. The resulting valid mask M is computed as the Iverson bracket:

$$M = \left[\frac{\|D_s - \gamma D_p\|^2}{\gamma D_p} < \sigma \right], \tag{6.12}$$

where σ is the threshold which we empirically set to 0.4 in our experiments, and γ is the alignment factor. For regularisation branch, γ is set to 1. For the pseudo labels created by self-supervised branch which are scale-ambiguous, we align them to the predictions from supervised branch by median scaling:

$$\gamma = \frac{median(D_s)}{median(D_p)}. \tag{6.13}$$

Then the pseudo labels are filtered to:

$$\hat{D}_p = \gamma M D_p. \tag{6.14}$$

6.3.5.2 *EMA Self-Training*

When updating the supervised branch, we adopt the EMA self-training scheme [TV17] for further performance gain and better robustness. To generate D_s, instead of directly using the updated supervised branch, we predict from a slow copy of the supervised model parameters θ_t at training step t by its exponential moving average over time, defined as:

$$\theta'_t = \alpha \theta'_{t-1} + (1 - \alpha)\theta_t, \tag{6.15}$$

where α is the smoothness factor which we empirically set to 0.99 in our experiments. For more details of EMA training technique please refer to [TV17].

6.3.5.3 *Adaptation Loss*

we update the supervised branch by computing SILog loss between its output D_s and the filtered pseudo labels created by the two branches:

$$\mathcal{L}_{ada} = \mathcal{L}_s(D_s, \hat{D}_{ps}) + \mathcal{L}_s(D_s, \hat{D}_{pr}), \tag{6.16}$$

where $\hat{D}_{ps}$ is the filtered pseudo label created by the supervised branch and $\hat{D}_{pr}$ is the one by the regularisation branch. $\mathcal{L}_s$ is the SILog loss defined in Equation 6.2.

6.4 Experiments

We perform extensive experiments and ablation study on various datasets to demonstrate the effectiveness of our framework. This section describes our experimental setup and demonstrates the results and discussions. More technical details of our implementation can be found in our code release.

6.4.1 Datasets

We conduct extensive experiments on various datasets on driving scenes and validate the effectiveness of our method.

KITTI dataset [GLSU13]. This dataset is captured in the city of Karlsruhe, Germany, including urban, rural and highway areas. The dataset includes 64 sequences of RGB images at 10fps with projected Lidar points as ground truth depth maps with a maximum range of 80m. Improved (denser) depth maps are provided on the KITTI depth estimation benchmark set [USS+17]. We adopt the same data splitting pattern as introduced in [EPF14], which splits the data into 32 seqs for training (∼40k frames) and 32 seqs for validation (∼4k frames). The camera height of this dataset is 1.65m and focal lengths are on average about 750mm.

DDAD dataset [GAP+20]. This dataset is recorded in several cities in the United States and Japan. It contains monocular videos at 10fps and the corresponding accurate ground-truth depth (across a full 360 degree field of view) generated from high-density LiDARs (up to 200m) mounted on a fleet of self-driving cars. The training set contains 150 scenes with a total of 12650 individual samples, and the validation set contains 50 scenes with a total of 3850 samples. The camera height is around 1.63m on average, and focal lengths around 2100mm.

Waymo dataset [ECC+21]. The more recent Waymo (perception) dataset is recorded across a range of conditions in multiple cities in the US, with large geographic coverage within each city. It comprises of images recorded by multiple high-resolution cameras and sensor readings from multiple high-quality LiDAR scanners (75m) mounted on a fleet of self-driving vehicles. It consists of around 1k videos, each about 200 frames at 10fps, under different weather conditions (sunny, rain) and time of day (day, dawn, night). The camera height is 2.12m, and focal lengths around 2000mm. In our experiments, we define several sub-sequences: **sunny-night-5** for the first 5 sequences of "sunny" night in the validation set; **rainy-5** for all of the 5 rainy scene from train set, including 1 during day, 3 in dawn and 1 at night; **all-6** for 6 sequences from the Waymo training set, including one for each time of day and weather ("day", "dawn" or "night"; "sunny" or "rain", all are the first ones from the category); **sunny-day-5** for the first 5 "sunny" "day" scene of the evaluation set.

Table 6.1: **Results on DDAD [GAP+20] validation set (cross-dataset).** We report the absolute scale results (without median scaling) of SOTA methods, SOTA + our scale alignment scheme (indicated as +SA), and our method. Best is illustrated in **bold**.

method	target ada	source sup	lower is better				higher is better		
			absRel	sqRel	RMSE	RMSE log	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
[XZH+20]	-	-	0.383	5.297	13.757	0.518	0.385	0.675	0.824
[YGD+22]	-	✓	0.437	8.211	18.498	0.817	0.249	0.362	0.480
[KPVG21]+sup	✓	✓	1.297	34.085	24.355	0.825	0.082	0.177	0.374
[WFVGD22]	✓	✓	0.441	8.313	18.598	0.829	0.246	0.358	0.473
[XZH+20]+SA	-	-	0.200	1.951	8.508	0.255	0.724	0.936	0.975
[YGD+22]+SA	-	✓	0.144	1.516	7.951	0.228	0.788	0.938	0.976
[XZH+20]+SA+Ada	✓	-	0.191	2.644	8.400	0.232	0.775	0.935	0.974
[KPVG21]+sup+SA	✓	✓	1.152	29.064	18.800	0.740	0.168	0.356	0.576
[WFVGD22]+SA	✓	✓	0.142	1.501	7.909	0.226	0.793	0.939	0.977
Ours	✓	✓	**0.112**	**1.173**	**6.649**	**0.186**	**0.867**	**0.961**	**0.984**

Table 6.2: **Results on the evaluation set of KITTI Eigen split [EPF14] (intra-dataset).** We report the results with and without ground truth median scaling ("scale factor" GT and -, respectively). Naïve adaptation counterparts of the methods indicated as "+ Ada". Best is illustrated in **bold**, second best in underline.

scale factor	method	target ada	source sup	lower is better				higher is better		
				absRel	sqRel	RMSE	RMSE log	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
GT	[GMAFB19]	-	-	0.115	0.903	4.863	0.193	0.877	0.959	0.981
	[XZH+20]	-	-	0.113	0.864	4.812	0.191	0.877	0.960	0.981
	[YGD+22] *	-	✓	**0.080**	0.445	3.476	0.155	0.928	**0.974**	**0.987**
	[GMAFB19]+Ada	✓	-	0.102	0.699	4.119	0.175	0.901	0.967	0.984
	[XZH+20]+Ada	✓	-	0.102	0.651	4.041	0.176	0.902	0.968	0.984
	[KPVG21]+sup	✓	✓	0.094	0.535	3.933	0.164	0.906	0.971	**0.987**
	[WFVGD22]	✓	✓	**0.080**	0.445	3.474	0.155	0.928	**0.974**	**0.987**
	Ours	✓	✓	0.081	**0.444**	**3.446**	**0.154**	**0.929**	**0.974**	**0.987**
-	[XZH+20]	-	-	0.118	0.925	4.918	0.199	0.862	0.953	0.979
	[YGD+22] *	-	✓	**0.077**	0.453	3.538	0.160	0.920	0.972	**0.986**
	[XZH+20]+Ada	✓	-	0.109	0.727	4.277	0.192	0.880	0.961	0.982
	[KPVG21]+sup	✓	✓	0.451	4.161	7.601	0.380	0.366	0.666	0.917
	[WFVGD22]	✓	✓	**0.077**	0.452	3.528	0.159	**0.921**	0.972	**0.986**
	Ours	✓	✓	0.078	**0.451**	**3.488**	**0.158**	**0.921**	**0.973**	**0.986**

* For fair comparison, we evaluate the methods using **raw Lidar points** as groundtruth, resulting in the numbers being different from those reported in the original paper [YGD+22] whose evaluations are done on the improved dense depth map.

6.4.2 Evaluation metrics

We evaluate depth predictions using the standard metrics described in [EPF14], which are computed by comparing the predicted depth and ground truth depth of the pixels where Lidar points are available. The error metrics, i.e. absolute relative difference (absRel), square relative difference (sqRel), root mean squared error (RMSE) and RMSE in logarithm space (RMSE log) are considered the lower the better, and for the 3 accuracy metrics where σ – the ratio of prediction and groundtruth – being under certain thresholds, higher is better. For more detailed formulation of these metrics please refer to [EPF14].

In our evaluation, we report both the results with and without the "groundtruth median scaling" operation described in [GMAFB19], where we scale the predictions with the ratio of the medians of groundtruth and predictions. (the same as described in Equation 6.13 and 6.14 where we conduct the operation between predictions and pseudo labels instead of groundtruth). The self-supervised models are scale-ambiguous so only with median scaling can the results

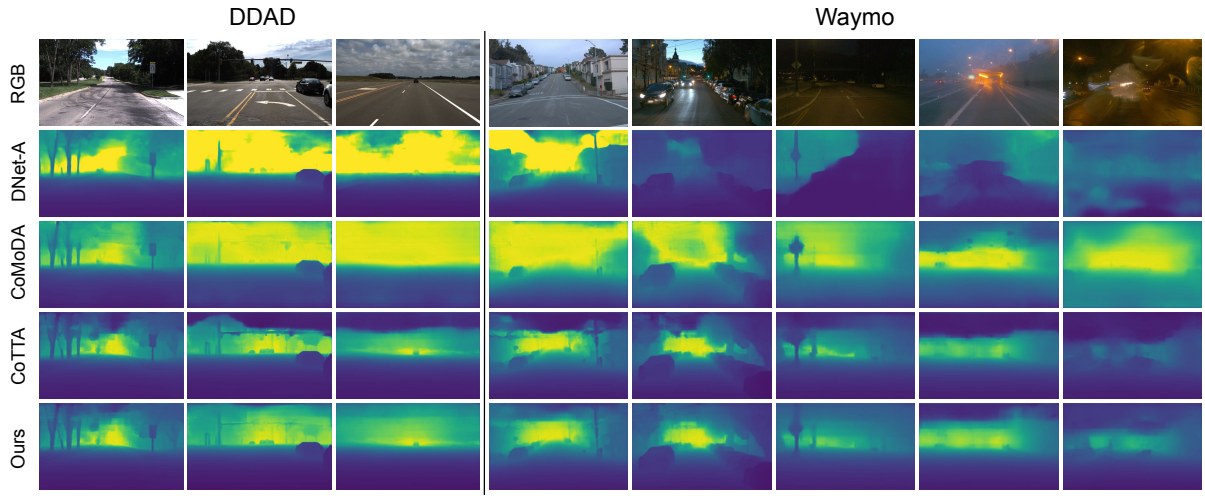


Figure 6.5: Qualitative results on DDAD and Waymo datasets.

Table 6.3: **Results on Waymo [ECC+21] dataset under different time of day/weather conditions (cross-dataset).** All methods are applied **after** our input scale alignment scheme, without median scaling. Best is illustrated in **bold**.

time/weather condition	method	target ada	source sup	lower is better				higher is better		
				absRel	sqRel	RMSE	RMSE log	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
clear night (sunny-night-5)	[XZH+20]	-	-	0.398	7.469	16.422	0.586	0.351	0.597	0.752
	[YGD+22]	-	✓	0.199	2.116	9.558	0.272	0.607	0.903	0.972
	[XZH+20]+Ada	✓	-	0.245	2.776	10.154	0.322	0.410	0.875	0.965
	[KPVG21]+sup	✓	✓	0.577	8.834	13.277	0.483	0.301	0.537	0.783
	[WFVGD22]	✓	✓	0.199	2.110	9.551	0.271	0.607	0.904	0.972
	Ours	✓	✓	**0.177**	**1.780**	**8.524**	**0.239**	**0.711**	**0.932**	**0.984**
rainy incl. day, dawn and night (rainy-5)	[XZH+20]	-	-	0.456	8.468	15.002	0.843	0.123	0.568	0.729
	[YGD+22]	-	✓	0.244	3.439	10.021	0.374	0.601	0.815	0.890
	[XZH+20]+Ada	✓	-	0.471	11.119	14.635	0.468	0.409	0.716	0.836
	[KPVG21]+sup	✓	✓	0.669	11.494	13.203	0.521	0.253	0.519	0.757
	[WFVGD22]	✓	✓	0.246	3.495	10.043	0.379	0.598	0.814	0.886
	Ours	✓	✓	**0.229**	**2.891**	**9.030**	**0.314**	**0.627**	**0.845**	**0.934**

be calculated; in real world practices, however, we expect the model to produce scale-aware predictions so we concern more about the results without groundtruth median scaling. Our method produces scale-aware depth maps, so the reported results "without median scaling" illustrates the general performance of our method and the ones "with median scaling" provide references to the model capacity of predicting the "local" information.

6.4.3 Experimental Results

We first evaluate our method on a cross-dataset setting, where we adapt the source models trained on the train set of KITTI Eigen split [EPF14] to the DDAD dataset to get absolute scale predictions, shown in Table 6.1. The DDAD dataset is recorded in daytime, in different cities with different camera setup compared to KITTI, thus resulting in a considerable domain shift. We observe that directly applying the state-of-the-art methods to the new dataset results in a huge performance drop, making the predictions nearly unreasonable. After applying our simple yet effective input scale alignment scheme, the performance improves on the state-of-the-art methods by a large margin; our method further improves significantly over the enhanced state-of-the-arts, including the naïve adaptation counterpart (noted as "+ Ada") of self-supervised methods ,

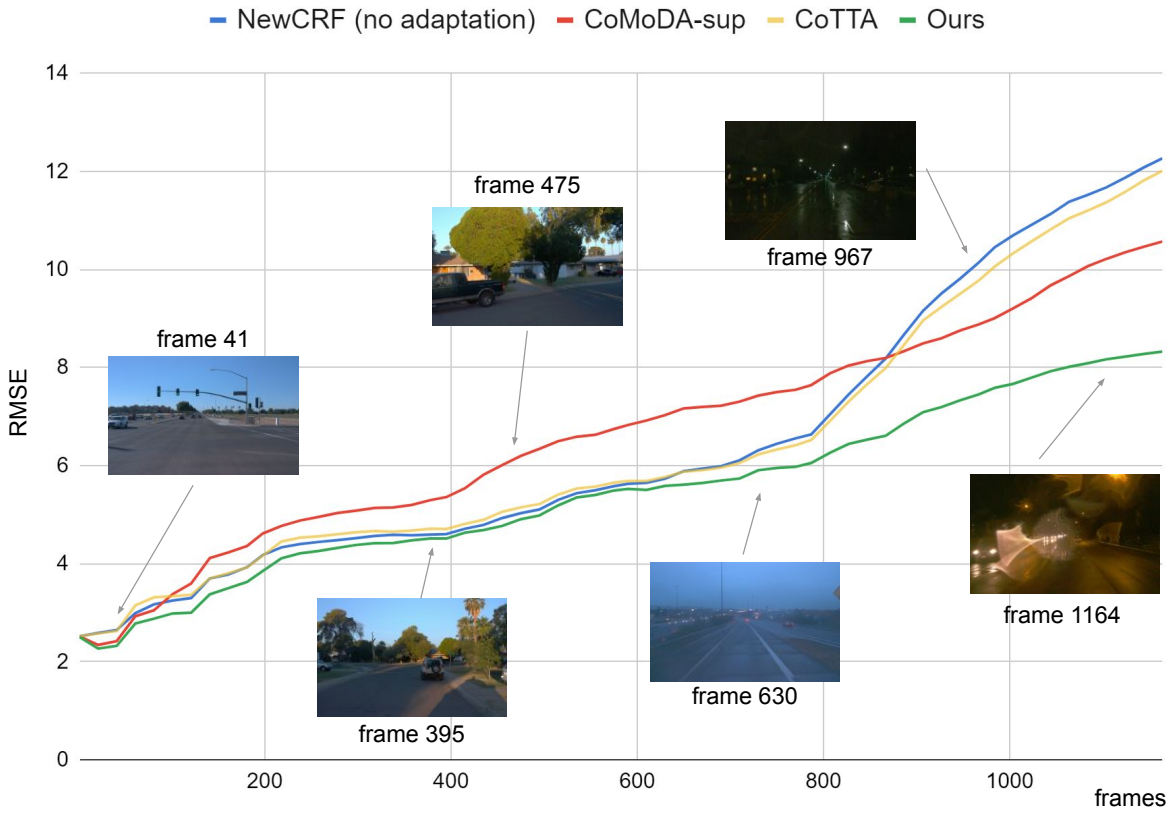


Figure 6.6: Accumulated average RMSE (averaged over all previous frames, lower is better) on Waymo "all-6".

which shows the effectiveness of both our easily-deployable input scale alignment scheme and our adaptation framework.

Compared to existing self-supervised baselines [GMAFB19, XZH+20] and domain adaptation methods [KPVG21, WFVGD22], our method is able to benefit from the Lidar groundtruth of the source data to produce significantly improved predictions on the target data, without requiring access to the source data or additional supervisory signal on the target data during adaptation. For fair comparison with the same data availability, we report the CoMoDA method [KPVG21] with source models trained with the Monodepth2 [GMAFB19] architecture on the KITTI training set with both supervised and self-supervised loss terms to get scale-aware depth and pose networks; then we apply the CoMoDA adaptation scheme on target data without "velocity supervision" (supervisory on test data) and "replay buffer" (access to source data) to align with our problem setting. We denote this variation as "CoMoDA[KPVG21]+sup". Note that our method continuously adapts to the entire DDAD validation set without suffering from catastrophic forgetting like [KPVG21]. We further observe that without our design components, the naïve adaptation on existing methods yields marginal improvements, but the performance is still far from being comparable to our method.

We also show intra-dataset results in Table 6.2. We adapt the source model trained on the KITTI train set and adapt to KITTI evaluation set. Our method still shows improvements on this setting, demostrating the effectiveness of our test-time adaptation scheme.

We further evaluate our method on a more challenging cross-domain setting by applying the

Table 6.4: **Ablative results on Waymo "sunny-day-5" (cross-dataset).** We report the results with and without ground truth median scaling ("scale factor" GT and -, respectively). Best is illustrated in **bold**, second best in underline.

scale factor	method	lower is better			
		absRel	sqRel	RMSE	RMSE log
GT	Monodepth2 (self-sup)	0.332	4.564	11.443	0.410
	+ scale align	0.229	4.697	11.314	0.277
	+ naïve adaptation	0.209	3.631	9.345	0.249
	NewCRF [YGD+22] (sup)	0.378	5.042	12.591	0.449
	+ scale align (focal)	0.186	2.106	8.014	0.237
	+ scale align (height)	0.157	1.615	6.853	0.210
	+ self-sup pseudo	0.196	2.988	8.422	0.233
	reg pseudo	0.157	1.599	6.802	0.209
	self-sup+reg pseudo	0.158	1.658	6.721	0.204
	+ pseudo consistency	0.154	1.644	6.825	0.205
	+ ema training	**0.147**	**1.590**	**6.652**	**0.196**
-	NewCRF [YGD+22] (sup)	0.482	9.234	19.413	0.858
	+ scale align (focal)	0.291	3.424	11.408	0.403
	+ scale align (height)	0.162	2.076	7.346	0.211
	+ self-sup pseudo	0.188	2.550	7.917	0.227
	reg pseudo	0.163	2.090	7.324	0.210
	self-sup+reg pseudo	0.163	2.063	**7.139**	0.205
	+ pseudo consistency	0.159	**2.056**	7.262	0.206
	+ ema training	**0.155**	2.130	7.241	**0.202**

KITTI trained source models on the Waymo [ECC+21] dataset, shown in Table 6.3. This dataset contains more diverse driving scenes in different cities, different time of day and weather conditions, with quite a different camera setup compared to that of the source data. To demonstrate the effectiveness of our method tackling large domain gaps, we evaluate our model on the defined "sunny-night-5" and "rainy-5" sequences. Our method shows significant improvements over existing methods and baselines (all without median scaling, after input scale alignment) in both cases.

We show some qualitative results in Fig. 6.5, adapting the KITTI-trained models on DDAD and Waymo dataset, respectively. Our method shows sharper and more plausible qualitative results, thanks to the information gain with our unsupervised test-time domain adaptation design.

Fig. 6.6 shows how our adaptation method behaves over time on challenging long, changing environments. We show the accumulated average RMSE (RMSE averaged over all previous frames) for the Waymo "all-6" sequences. Our method adapts better to the new changes with less drift.

6.4.4 Ablation Study

We perform ablative study on the Waymo "sunny-day-5" sequences, shown in Table 6.4. We incrementally add every component to the baseline (the supervised branch of our method, based

on [YGD+22]). The results show that every design choice helps with enhancing the performance, and our full model acquires the best performance.

6.5 Conclusion

We propose a source-free, online test time domain adaptation method for monocular depth estimation. Our input scale alignment scheme significantly improves the network inference in the presence of large domain gaps even before the adaptation, and can be easily integrated into any deep learning based depth estimation framework. Our 3-branch self-training framework shows superiority over the naïve adaptation of either supervised or self-supervised framework. Our effective regularisation operation keeps the learning stable while not degrading the performance gain. Extensive experiments show that our method achieve state-of-the-art results in various datasets, especially in the challenging scenario of large domain gap and continuously changing environment. The practical setting enables our method to be applied to data-scarce, low-cost systems. An interesting future direction can be to reduce the model sizes and/or lighten the training computations in order to deploy it in real-time applications.

7 3D Occupancy Prediction via Multi-Task Distillation

Contents

3D occupancy prediction is gaining traction in autonomous driving for its ability to jointly model environment geometry and semantics. Weakly supervised methods learn 3D representations solely from multi-view 2D labels, making them ideal for data-scarce scenarios. However, distilling 2D knowledge into 3D is challenging due to limited information and noisy pseudo labels. To tackle the above challenges, we introduce MT-Occ, a single-view self-supervised 3D occupancy prediction method that enhances 2D-to-3D distillation by leveraging pretrained multi-task features and modelling task interactions. Our approach includes three effective and flexible components: 1) an effective fusion technique utilising pretrained features of multiple relevant tasks, 2) a spatial cross-task attention module for geometric-semantic distillation and 3) a view-consistent label refinement strategy to improve 2D pseudo labels. MT-Occ achieves state-of-the-art results on autonomous driving benchmarks, outperforming prior work with +16.04% relative mIoU on SSCBench-KITTI-360 and +33.33% on SSCBench-nuScenes, particularly excelling on safety-critical small classes. Extensive experiments validate the effectiveness and flexibility of our design choices.

This chapter is based on [LARS25]. As the first author, Zhi Li worked on all parts of the project, including idea generation, all implementation and experiments, and paper writing.

7.1 Introduction

Accurate 3D scene understanding is crucial for building reliable autonomous driving systems. The task of 3D occupancy prediction [Mob20, Tes21, TSW+23, WZX+23] proposed in recent years leverages 2D images to capture rich 3D spatial and semantic information, providing critical scene context for autonomous driving perception and planning while reducing the reliance on costly LiDAR sensors. However, training fully supervised 3D occupancy prediction networks requires accurate 3D ground truth (e.g., dense LiDAR data with semantic labels), which is costly and error-prone to collect at scale. Additionally, many existing methods require multi-view

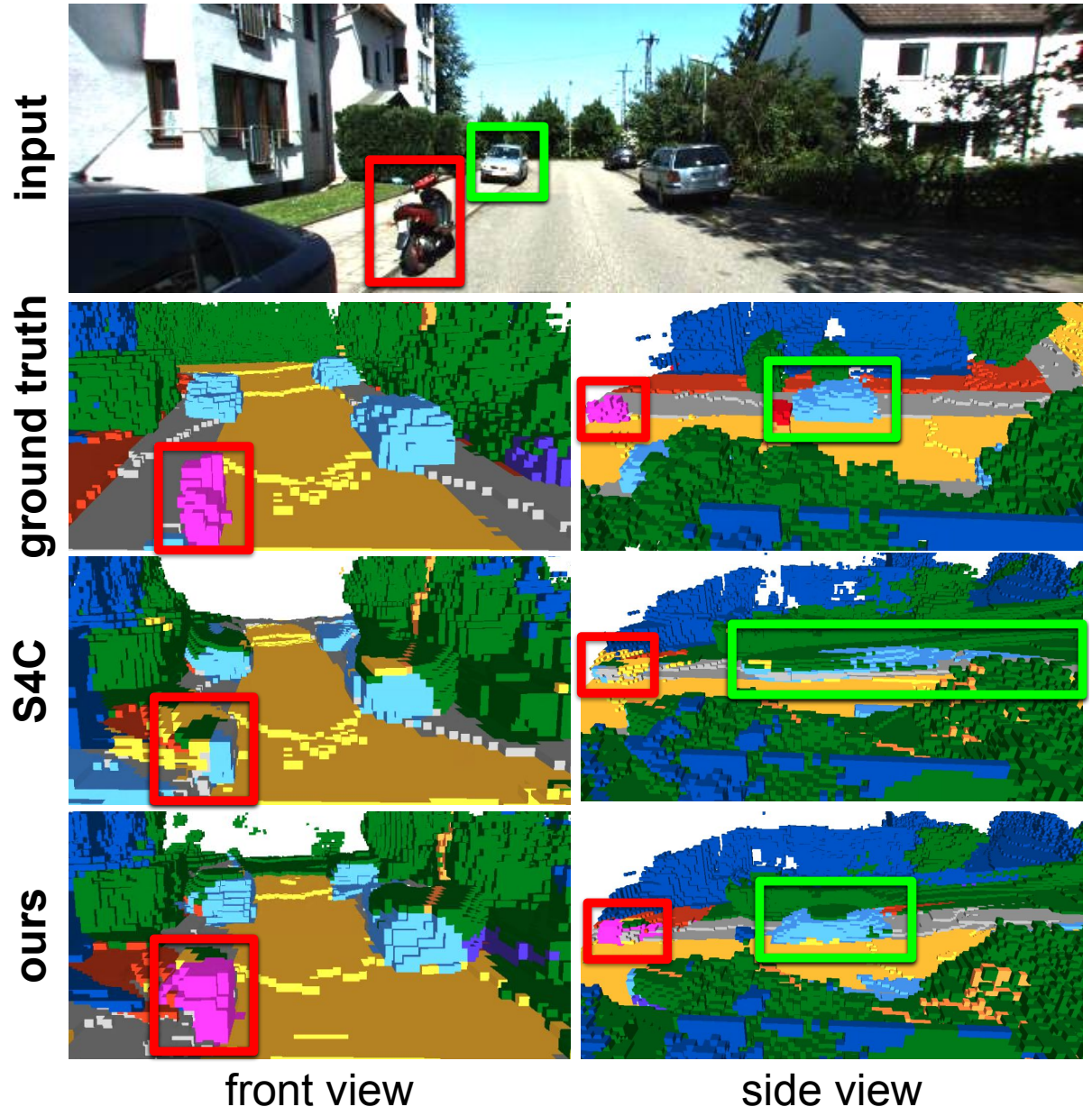


Figure 7.1: **Single-view 3D occupancy prediction results.** Given a single-view image as input (top), we show the 3D occupancy results of the existing method S4C [HWM+24] and our MT-Occ, from a front viewpoint (left) and a side viewpoint (right). Our method can capture objects of challenging classes such as motorcycle (highlighted in red boxes), and reconstructs more faithful geometry with semantics especially on self-occluded areas (shown in green boxes).

or multi-frame images as inputs during inference, which limits flexibility and increases system costs by necessitating specific camera setups and calibration processes [TSW+23, WZX+23, LYW+23, LWL+22, PLZ+24, ZDM23]. To develop more adaptable and affordable systems, it is advantageous to use approaches that can be trained with only 2D labels and require only single-view inputs for inference. However, this setup amplifies the inherent difficulty of 3D occupancy estimation, as it demands robust self-supervised methods to compensate for the lack of explicit 3D information and to address the ill-posed nature of the single-view task.

Existing approaches [HWM+24, ZYW+23, HZZ+24a, PLZ+24] generally achieve self-supervision of 3D occupancy prediction by distilling multi-view 2D information into 3D with neural rendering [MST+21, YYTK21], in which continuous neural density fields together with semantic fields are learned from posed images and 2D semantic (pseudo) labels, avoiding the need for 3D groundtruth. However, transferring knowledge from 2D to 3D can be constrained by an information bottleneck, as 2D representations have limited capacity to fully express 3D structures, and the noise in 2D pseudo labels further adds to the challenge. As shown in Fig. 7.1, existing methods, whose distillation relies on simple learning from limited and noisy 2D pseudo labels, often fail to capture some certain small "thing" classes (e.g. bicycle, motorcycle, person) which are safety-critical but challenging and appear rarer in the training set, and demonstrate unfaithful geometry reconstruction on the side view (termed in previous works [WYRC23, HWM+24, LFS+24] as "trailing artifacts" as the objects tend to "trail" towards the camera ray direction in self-occluded areas).

To address the 2D to 3D data distillation challenge, we argue that the following 3 aspects are essential: enriching 2D feature representation, improving information transfer during network training, and reducing noise in 2D pseudo labels. Recent vision foundation models [ODM+25, RKH+21, KMR+23, YKH+24a, YKH+24b] trained on massive datasets offer informative general feature representations that enhance 2D tasks. Although not designed for 3D tasks, these pretrained models can help improve 3D occupancy prediction by refining feature representations and pseudo-labels for tasks like semantic segmentation and depth estimation. Additionally, geometry and semantics in 3D occupancy prediction are tightly coupled, where one can aid the other [XOWS18, VVGG+20, XZV+22], which can be utilised for both network training and pseudo-label regularisation.

Here, we propose MT-Occ, a new method to improve single-view, self-supervised 3D occupancy prediction by distilling information from relevant 2D pretraining and multi-task interactions. Our key contributions includes:

- An effective feature fusion technique for single-view 3D occupancy prediction that leverages pretraining from relevant 2D tasks.
- A spatial cross-task attention mechanism in the decoding phase enhancing geometry-semantics interaction.
- A novel label refinement strategy, using relative depth estimation as a proxy to guide 3D occupancy training and refine noisy pseudo-labels.
- Extensive experiments on SSCBench-KITTI-360 and SSCBench-nuScenes benchmarks demonstrate the effectiveness and flexibility of all proposed components

7.2 Related Works

Semantic scene completion and 3D occupancy prediction. Semantic Scene Completion (SSC) [SYZ+17, RDCVB22, LLL+24] integrates 3D geometry and semantic labelling in a voxelised 3D space, using either images together with partial geometry [CAR+21, LHW+20, LLG+19, LLY+19, LHZ+18, WTNT20, RdCVB20, YGL+21, REEG21] or solely from RGB images [CDC22, YLS+23, HWM+24, CDdC24]. Recently, "3D occupancy prediction" has gained attention in autonomous driving to refer specifically to SSC methods using vision-only inputs [TSW+23, WZX+23]. While some methods use multi-view, multi-frame inputs [TSW+23, WZX+23, PLZ+24, WZZ+23, HZZ+24b, HTZ+25], others focus on single-view (monocular,

single-frame) input [CDC22, HWM+24], which simplifies setup but increases reconstruction difficulty. We target this challenging single-view 3D occupancy prediction, proposing effective strategies for 2D-to-3D inference.

Self-supervised 3D occupancy prediction. Methods trained only with 2D (pseudo) labels [HWM+24, ZYW+23, HZZ+24a, PLZ+24] are dubbed in the previous literatures as self-supervised methods. They typically learn a density-semantic NeRF [MST+21] from posed multi-view images and 2D semantic labels through neural rendering. Most methods [ZYW+23, PLZ+24, HZZ+24a] convert multi-view inputs into BEV features, requiring specific multi-camera setups. Some recent methods [GLX+24, CZB+25, BGR25, ZWW+25] leverage Gaussian representation [KKLD23] to build multi-view occupancy frameworks. A recent approach [HWM+24] tackles single-view, monocular input for practical 3D occupancy, requiring multi-view data only during training. We build on this setup, introducing a single-frame framework that improves performance via multi-task distillation.

Depth and semantic priors for NeRFs. Several works demonstrate that leveraging depth estimation or semantic information improves NeRF training. Depth estimation is used in various designs to guide or supervise neural rendering for accuracy or efficiency improvements [YPN+22, YHL+23, PHT23, WLZ+24, CHL+24]. Additionally, 2D semantic segmentation or panoptic segmentation distills semantic information into NeRF [ZLLD21, FZC+22, KGY+22], and vision-language features similarly enhance scene understanding [FYJ+22, KKG+23, PGJ+23, LFS+24]. While these works focus on generating novel views, segmenting objects, or scene reconstruction, we develop techniques specifically for 3D occupancy prediction by distilling multi-task priors.

Multi-task learning. Jointly predicting multiple 2D tasks from input images is a problem setup addressed by the concept of multi-task learning [XOWS18, VVGG+20, FAZ+21, LLJ+21, BMAZ22, BZSS22, LVdC23, XZV+22], in which the multi-task interactions, including depth estimation and semantic segmentation, are modelled by various distillation designs enhance one or more tasks in the frameworks. While these approaches focus on 2D tasks, we extend multi-task interaction to 3D occupancy prediction, aiming to improve joint 3D scene reconstruction and semantic prediction.

7.3 Method

Our method focuses on the challenging setting of 3D occupancy prediction from a single RGB image without 3D supervision. We illustrate the proposed MT-Occ framework in Fig. 7.2. In this section, we first introduce our base framework that achieves single-view 2D-supervised 3D occupancy prediction (Sec. 7.3.1), then detail our proposed components for multi-task feature fusion (Sec. 7.3.2), spatial cross-task attention (Sec. 7.3.3) and view-consistent label refinement (Sec 7.3.4).

7.3.1 Framework Overview

MT-Occ reconstructs the full 3D scene from single-view inputs through a generalisable NeRF [YYTK21] based encoder-decoder network. Given a single RGB image $I_0 \in \mathbb{R}^{3\times H\times W}$, along with its corresponding camera intrinsics $K_0 \in \mathbb{R}^{3\times 4}$ and extrinsics $T_0 \in \mathbb{R}^{4\times 4}$, the network encodes the full 3D scene into a dense, pixel-aligned, implicit, and continuous feature field $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$. This feature field represents the density and semantic distributions along the rays cast from the optical centre through the pixels. Given a 3D point $\mathbf{x} \in \mathbb{R}^3$ in the world coordinate system, the density-semantic field $\mathbf{F}$ can be queried at the point's projected location $\mathbf{u} = \pi_0(\mathbf{x})$ on the

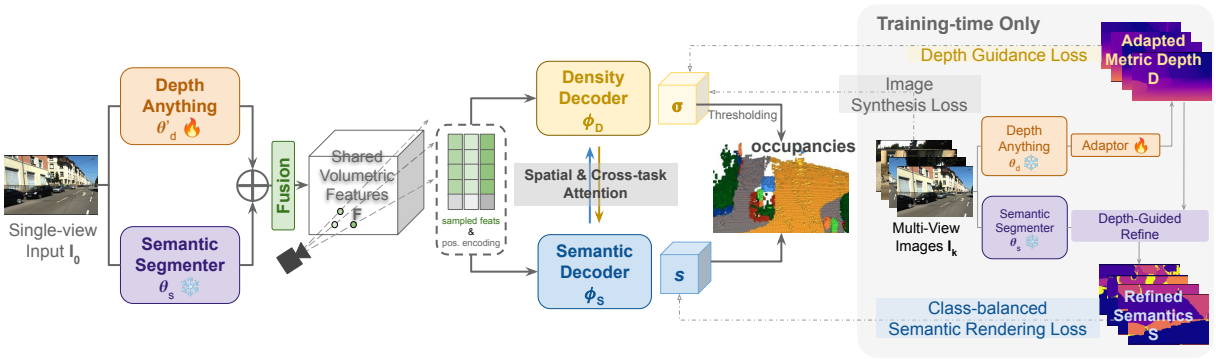


Figure 7.2: **MT-Occ framework.** Given single-view image inputs, we leverage pretrained depth and semantic networks to extract pixel-aligned implicit features describing the density and semantic distributions along rays cast through each pixel. These features, concatenated with positional encodings of 3D points, are decoded into density and semantic values using decoders with the proposed spatial cross-task attention modules, then aggregated for the final 3D occupancy prediction. During training, multi-view pseudo-2D semantic labels for the rendering loss are refined via depth estimation for view consistency, while the adapted depth maps guide density training.

2D pixel plane, determined by the projection operation $\pi_0(\mathbf{x}) = K_0 T_0 \mathbf{x}$. The extracted feature $f_\mathbf{u}$ is then concatenated with the positional encoding $\gamma(\mathbf{x})$ and decoded by a density decoder $\phi_D(\cdot)$ and a semantic decoder $\phi_S(\cdot)$ to obtain the density value $\sigma \in [0, 1]$ and the semantic class $s \in \{0, ..., c-1\}$, where c is the total number of classes. A threshold τ is applied to σ to obtain a binary occupancy representation. Throughout the encoding-decoding process, MT-Occ enhances feature extraction in $\mathbf{F}$ through multi-task feature fusion (Sec. 7.3.2) and improves density and semantic decoding using a spatial cross-task attention module in the decoders $\phi_D(\cdot)$ and $\phi_S(\cdot)$ (Sec. 7.3.3).

During training, multi-view images and their corresponding (pseudo) semantic maps supervise the NeRF-rendered RGB images and semantics, eliminating the need for ground truth 3D labels. To render a viewpoint from the semantic-density field, rays are cast from the camera through each pixel. The rendered colour or semantic value of a pixel is computed as the weighted integral of the colour or semantic values of 3D points $\mathbf{x}$ along the ray, using the points' probabilities T of not being occluded. In practice, the integral is approximated by the weighted sum of points $\{\mathbf{x}_i | i \in \{0, ..., m\}\}$ at m discrete steps along the ray. Specifically, for the step i, the probability T_i of $\mathbf{x}_i$ being not occluded is given by

$$\alpha_i = exp(1 - \sigma_{\mathbf{x}_i} \delta_i), \qquad T_i = \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{7.1}$$

where δ_i denotes the distance between adjacent sampled points $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$, and $\sigma_{\mathbf{x}_i}$ the density prediction at $\mathbf{x}_i$.

To render the colour values $\hat{c}$, we aggregate the colour values sampled from other view images $\mathbf{I}_k$ by projecting the 3D points $\mathbf{x}$ along the ray onto these views, yielding $c_{\mathbf{x},k} = \mathbf{I}_k(\pi_k(\mathbf{x}))$. The rendered colour of the pixel (w.r.t. the source view k) is then given by

$$\hat{c}_k = \sum_{i=1}^{m} T_i \alpha_i c_{\mathbf{x}_i,k}. \tag{7.2}$$

Similarly, the semantic class s of the pixel is rendered by aggregating all predicted semantic logits s of the 3D points along the ray:

$$\hat{s} = \arg\max \sum_{i=1}^{m} T_i \alpha_i \cdot \text{softmax}(\text{s}_{\mathbf{x}_i}). \tag{7.3}$$

Additionally, we can render the depth value of the pixel by retrieving the expected ray termination depth $\hat{d}$:

$$\hat{d} = \sum_{i=1}^{m} T_i \alpha_i d_i \tag{7.4}$$

The network training can then be supervised by calculating the losses between the rendered colour, semantics and depth and the corresponding (pseudo) groundtruth 2D labels. At training stage, MT-Occ models multi-task interactions, proposes strategies for view consistent pseudo-label refinement and enables more robust training (Sec. 7.3.4).

7.3.2 Multi-Task Feature Fusion

At the feature extraction stage, the occupancy network converts single-view RGB inputs into pixel-aligned volumetric features, encoding scene density and semantics within the camera frustum. Prior methods [CDC22, HWM+24] achieve feature extraction by training standard architectures (e.g., ResNets [HZRS16], UNets [RFB15]) only within the scope of the limited training data. However, this approach is less effective for complex tasks like 3D occupancy prediction, where targeted features are crucial and extensive data is often needed for effective learning.

In NeRF based frameworks, extracted features describe density and semantic distributions along rays extending from the camera centre through each pixel. Intuitively, dense 2D perception tasks like depth estimation and semantic segmentation can aid this learning. To leverage depth priors, we initialise our feature extractor with pretrained weights θ_d from DepthAnythingv2 [YKH+24b], a foundation model trained on large-scale synthetic data. We modify θ_d by removing its final convolutional layer, yielding θ'_d, which retains its last-layer feature map but excludes direct depth predictions. Since depth estimation encodes only partial geometry (i.e., surfaces), we update θ'_d during training to transform depth information into a density distribution along a ray, capturing full 3D geometry, including occluded regions. Given an input image $\mathbf{I}_0$, we then generate the feature representation of density distribution $\mathbf{F}_d$ by

$$\mathbf{F}_d = f(\mathbf{I}_0, \theta'_d). \tag{7.5}$$

We enhance the feature representation with semantic information using predictions from a frozen off-the-shelf semantic segmentation network θ_s. To ensure flexibility across architectures, we use its one-hot encoded output $\mathbf{F}_s$:

$$\mathbf{F}_s = f(\mathbf{I}_0, \theta_s)^{onehot}. \tag{7.6}$$

Next, we concatenate $\mathbf{F}_d$ and $\mathbf{F}_s$, then apply a convolutional layer for fusion, yielding $\mathbf{F}$, a feature representation enriched with both geometry and semantics information:

$$\mathbf{F} = \text{conv}(\text{concat}(\mathbf{F}_d, \mathbf{F}_s)). \tag{7.7}$$

The fused feature $\mathbf{F}$ describes a density-semantic field of the scene, and can then go through the decoding process to predict density and semantic values of 3D points.

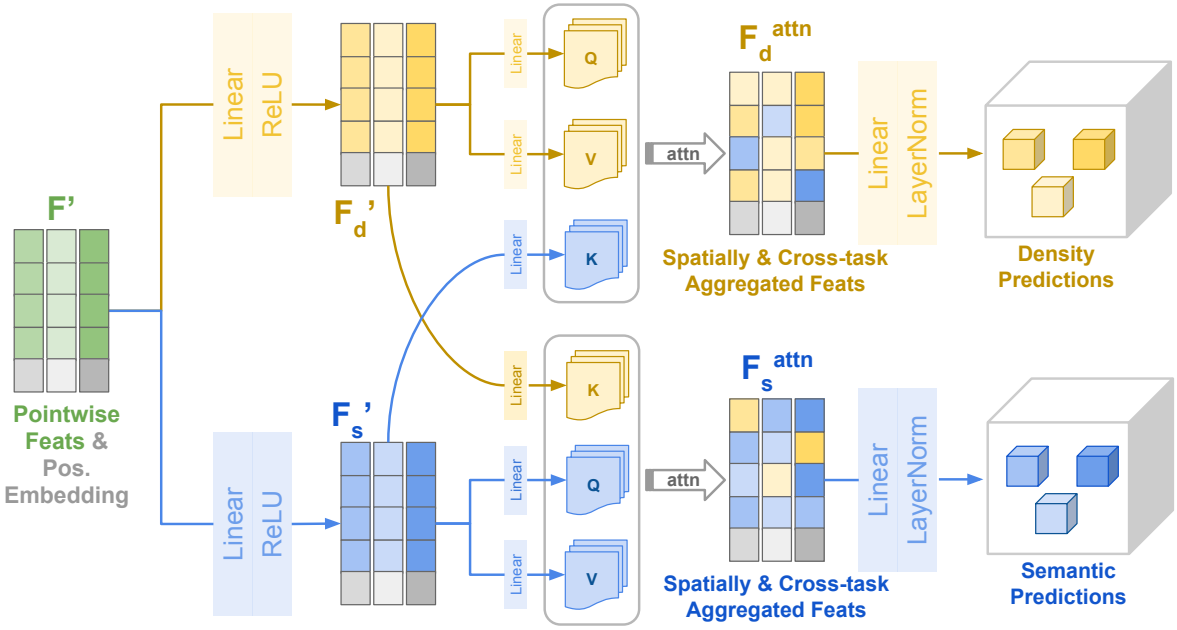


Figure 7.3: **Spatial cross-task attention module.** We aggregate features of different 3D points and calculate cross-task attention from different task branches with a linear attention design.

7.3.3 Spatial Cross-Task Attention

During decoding, sampled features from $\mathbf{F}$ are concatenated with positional embeddings and fed into task decoders ϕ_d and ϕ_s for density and semantic predictions. Unlike prior methods [HWM+24, PLZ+24, ZYW+23, HZZ+24a] that use separate MLP heads per 3D point, we follow [LFS+24] to aggregate features from neighbouring points, enhancing geometry learning. Additionally, we introduce a cross-task attention module to facilitate interaction between geometry and semantics (Fig. 7.3).

Given sampled and positionally embedded features $\mathbf{F}'$, task-specific features $\mathbf{F}'_d$ and $\mathbf{F}'_s$ are computed via fully connected layers. We then apply an efficient linear attention mechanism [SZZ+21, LFS+24], using features from the same task as query $\mathbf{Q}$ and value $\mathbf{V}$, while the other task serves as key $\mathbf{K}$. The global context score G is computed by attending to $\mathbf{K}$ and $\mathbf{V}$, then correlated with $\mathbf{Q}$ to obtain the final attended features $\mathbf{F}_d^{\text{attn}}$ and $\mathbf{F}_s^{\text{attn}}$. Specifically, for density, we have

$$\begin{aligned} G_d &= \text{softmax}(\mathbf{K}_s/\sqrt{D})^\top \cdot \mathbf{V}_d, \\ \mathbf{F}_d^{\text{attn}} &= \text{softmax}(\mathbf{Q}_d/\sqrt{D}) \cdot G_d, \end{aligned} \tag{7.8}$$

and similarly for semantics

$$\begin{aligned} G_s &= \text{softmax}(\mathbf{K}_d/\sqrt{D})^\top \cdot \mathbf{V}_s, \\ \mathbf{F}_s^{\text{attn}} &= \text{softmax}(\mathbf{Q}_s/\sqrt{D}) \cdot G_s, \end{aligned} \tag{7.9}$$

with D being the scaling factor. The spatially and cross-task aggregated features $\mathbf{F}_d^{\text{attn}}$ and $\mathbf{F}_s^{\text{attn}}$ is then fed into the last layers for the final density and semantic prediction.

7.3.4 View-Consistent Label Refinement

The self-supervised 3D occupancy prediction network is trained using only 2D supervisory signals. During training, it extracts a feature map $\mathbf{F}$ from a single-view input $\mathbf{I}_0$, while multi-view images $\mathbf{I}_k$ (aggregated from the main, stereo, and side-view cameras over multiple video frames, as in [HWM+24]) provide supervision. Prior work [HWM+24] employs a photometric discrepancy loss $\mathcal{L}_{\text{ph}}$, combining L1 and structural similarity (SSIM [WBSS04]), computed between randomly sampled patches $\mathbf{P}$ (from $\mathbf{I}_0$) and reconstructed patches $\hat{\mathbf{P}}_k$ (rendered from $\mathbf{I}_k$) using Eq. 7.2,

$$\mathcal{L}_{\text{ph}} = \min_{k \in N_{\text{render}}} (\text{L1}(\mathbf{P}, \hat{\mathbf{P}}_k) + \lambda \text{SSIM}(\mathbf{P}, \hat{\mathbf{P}}_k)), \tag{7.10}$$

and a regularisation loss, dubbed as edge-aware smoothness loss, applied on inverse, mean-normalised reconstructed depths $\mathbf{d}^* = \hat{\mathbf{d}}/\bar{\hat{\mathbf{d}}}$ ($\hat{\mathbf{d}}$ are reconstructed depths from Eq. 7.4) proposed by [GMAFB19],

$$\mathcal{L}_{\text{eas}} = |\partial_x \mathbf{d}^*| \, e^{-|\partial_x \mathbf{P}|} + |\partial_y \mathbf{d}^*| \, e^{-|\partial_y \mathbf{P}|}, \tag{7.11}$$

where $\partial_x(\cdot)$, $\partial_y(\cdot)$ are partial derivatives along the x, y axis.

For semantics, a binary cross-entropy loss $\mathcal{L}_{sem}$ is applied between sampled patches of 2D pseudo-semantic labels $\mathbf{S}$ (predicted by off-the-shelf semantic networks) and reconstructed semantic patches $\hat{\mathbf{S}}$ (computed via Eq. 7.3):

$$\mathcal{L}_{\text{sem}} = \text{BCE}(\mathbf{S}, \hat{\mathbf{S}}). \tag{7.12}$$

However, direct predictions $\mathbf{S}$ from off-the-shelf semantic networks are often noisy. To refine the semantic pseudo-labels $\mathbf{S}$ during training, we leverage multi-view consistency with the aid of depth estimation. Specifically, given a relative depth estimate from DepthAnythingv2 [YKH+24b] for a target image $\mathbf{I}_{\text{target}}$, we first apply two small convolutional layers as an adaptor to convert the relative depth into absolute-scale depth $\mathbf{D}_{\text{target}}$. These adaptors are trained in a self-supervised manner using the photometric reprojection loss from [GMAFB19], which measures the discrepancy between $\mathbf{I}_{\text{target}}$ and its reconstructed version $\hat{\mathbf{I}}_{\text{target}}$. The reconstruction is achieved by warping pixels from source images $\mathbf{I}_{\text{source}}$ to the target view using $\mathbf{D}_{\text{target}}$ and camera parameters, following the formulation in [GMAFB19]. With the learned absolute-scale depth map $\mathbf{D}_{\text{target}}$, we apply the same reconstruction process to refine the semantic pseudo-labels. Specifically, we reconstruct $\hat{\mathbf{S}}_{\text{target}}$ from $\mathbf{S}_{\text{source}}$ and consider only the pixels $\mathbf{u}'$ whose reconstructed labels match the original pseudo-labels as reliable, filtering out the rest, given by

$$\mathbf{u}' = [u | \mathbf{S}_{\text{target}}(u) = \hat{\mathbf{S}}_{\text{target}}(u)]. \tag{7.13}$$

In addition, we propose semantic-guided class-balanced patch sampling to stabilise training on class-imbalanced datasets. Unlike previous random sampling methods, which bias toward larger or more frequent classes, we enforce equal sampling probability for all classes in $\mathbf{S}_{\text{target}}$, yielding class-balanced patches $\mathbf{S}'$. Thus, Eq. 7.12 becomes

$$\mathcal{L}_{\text{sem\_refine}} = \text{BCE}(\mathbf{S}'(\mathbf{u}'), \hat{\mathbf{S}}'(\mathbf{u}')). \tag{7.14}$$

We further apply a depth supervision loss using our refined depth map $\mathbf{D}$ on the rendered depth map $\hat{\mathbf{D}}$ (via Eq. 7.4). We adopt the scale-invariant SILog loss proposed by [EPF14]. Let n be the number of pixels in $\mathbf{D}$ and $d_j, \hat{d}_j$ the depth value of a pixel j in $\mathbf{D}, \hat{\mathbf{D}}$

$$\mathcal{L}_{\text{depth}} = \frac{1}{n} \sum_i a_j^2 - \frac{1}{n^2} (\sum_i a_j)^2, a_j = \log \hat{d}_j - \log d_j. \tag{7.15}$$

Our final loss function $\mathcal{L}$ is given by

$$\mathcal{L} = \mathcal{L}_{\text{ph}} + \lambda_{\text{e}}\mathcal{L}_{\text{eas}} + \lambda_{\text{s}}\mathcal{L}_{\text{sem\_refine}} + \lambda_{\text{d}}\mathcal{L}_{\text{depth}}. \tag{7.16}$$

7.4 Experiments

7.4.1 Experimental Setup

Datasets. We evaluate our method on SSCBench-KITTI-360 and SSCBench-nuScenes datasets [LLL+24]. **SSCBench-KITTI-360** is a subset of KITTI-360 [LXG22], containing 80% of the data with multi-view sequences from forward-facing stereo and fisheye side cameras. Following [HWM+24], we sample 2 frames per view within 4s (8 views per sample) for training. Ground-truth semantic occupancy is derived from aggregated LiDAR annotations every 5 frames, with the dataset comprising 42k training frames (7 sequences), 15k validation frames (1 sequence), and 13k test frames (1 sequence), totaling 2566 test frames. **SSCBench-nuScenes** is derived from nuScenes [CBL+20], providing similar 3D occupancy ground truth but aligned to single-view front-facing sequences. The original nuScenes dataset includes six synchronized surround-view cameras, enabling multi-view training. We sample 4 frames from the front camera and 4 from randomly chosen front/back left/right cameras. The dataset consists of 850 20-second scenes, split into 500 scenes (∼20k frames) for training, 200 (∼8k) for validation, and 150 (∼6k) for testing, with ground-truth voxels available for all frames.

Evaluation. We follow the standard setup [REEG21, LLL+24, TSW+23, WZX+23, CDC22, LYC+23, ZZD23, HZZ+23, HWM+24] to evaluate scenes of size $51.2m \times 51.2m \times 6.4m$ at a $0.2m$ voxel resolution. Using the threshold $\tau = 0.1$ from [HWM+24] whose selection is based on validation set AuC, we convert predicted densities to binary occupancy. We report IoU (intersection over union) against the 3D voxel groundtruth to assess geometric reconstruction quality and mIoU (mean intersection over union) together with per-class IoUs for semantic reconstruction quality. We further report mIoU and per-class IoU for rendered 2D semantic maps against 2D groundtruth provided for the front views in SSCBench-KITTI-360, to access the front view rendering quality.

	Method	IoU (%)	mIoU (%)	car (2.85%)	bicycle (0.01%)	motocycle (0.01%)	truck (0.16%)	other-veh. (5.75%)	person (0.02%)	road (14.98%)	sidewalk (6.43%)	building (15.67%)	fence (0.96%)	vegetation (41.99%)	terrain (7.10%)	pole (0.22%)	traf.-sign (0.06%)	other-obj. (0.28%)
3d-gt-sup	[CDC22]	37.87	13.52	19.34	0.43	0.58	8.02	2.03	0.86	48.35	28.13	32.89	3.53	26.15	16.75	6.92	5.67	3.09
	[LYC+23]	38.76	13.20	17.84	1.16	0.89	4.56	2.06	1.63	47.01	27.21	31.18	4.97	28.99	14.69	6.51	6.92	2.43
	[HZZ+23]	40.22	14.95	21.56	1.09	1.37	8.06	2.57	2.38	52.99	31.07	34.83	4.80	30.08	17.51	7.46	5.86	2.70
	[ZZD23]	**40.27**	14.97	**22.58**	0.66	0.26	9.89	3.82	2.77	**54.30**	31.53	**36.42**	4.80	**31.00**	**19.51**	**7.77**	**8.51**	**4.60**
	[HZZ+24b]	35.38	13.79	18.93	1.02	**4.62**	**18.07**	**7.59**	**3.35**	45.47	25.03	28.44	**5.68**	29.54	8.62	2.99	2.32	5.14
	[HTZ+25]	38.37	**15.33**	21.08	**2.55**	4.21	12.41	5.73	1.59	54.12	**32.31**	32.01	4.98	28.94	17.33	3.57	5.48	3.54
2d-psu	[HWM+24]	38.84	10.10	10.32	0	0	2.17	0.15	**0.36**	48.61	26.43	20.96	2.80	22.32	16.45	0.43	**0.49**	0
	MT-Occ	**40.44**	**11.72**	**12.75**	**2.48**	**4.17**	**6.90**	**2.74**	0.34	**52.26**	**26.80**	**22.84**	**3.57**	**23.15**	**16.53**	**0.49**	0.34	**0.08**

Table 7.1: **Single-view 3D occupancy prediction results on SSCBench-KITTI-360.** Our method achieves state-of-art performance against existing 2D pseudo-supervised single-view method (indicated as **2d-psu**), and even surpassing 3D groundtruth supervised methods (**3d-gt-sup**) on some rare "thing" classes. (%) indicates each class label ratio. Best numbers in **bold**.

	Method	IoU (%)	mIoU (%)	car (2.47%)	bicycle (0.16%)	motocycle (0.02%)	truck (0.96%)	other-veh. (7.87%)	person (0.24%)	road (36.97%)	sidewalk (8.30%)	building (21.17%)	vegetation (20.97%)	other-obj. (0.11%)
3d-gt-sup	[CDC22]	**29.63**	9.34	10.17	1.70	3.80	8.35	8.74	3.72	**38.77**	14.74	7.23	5.50	0.03
	[LYC+23]	25.16	5.04	4.95	0.29	1.21	2.73	2.45	1.12	23.94	10.14	3.97	4.58	**0.06**
	[ZZD23]	28.23	**11.24**	**14.61**	**2.25**	**7.97**	**11.88**	**9.80**	**5.87**	37.62	**18.63**	**9.05**	**5.92**	0
2d-psu	[HWM+24] - nuscs	17.98	4.20	4.27	0	0	0	0	1.07	21.74	8.88	4.94	5.14	0.14
	MT-Occ - nuscs	22.84	5.29	5.38	0.64	0.63	0	0	1.21	28.35	9.12	5.64	7.01	0.15
	[HWM+24] - kt360	12.44	3.37	1.81	0	0	0	0	0.73	23.36	4.99	2.49	3.70	0.14
	MT-Occ - kt360	15.79	4.63	2.10	0.59	1.16	0	0	1.26	32.58	5.33	2.97	4.97	0.07
	[HWM+24] - ft	18.24	5.07	6.99	0	0	0	0	1.80	26.09	10.03	5.09	5.81	0.14
	MT-Occ - ft	**24.50**	**6.76**	**7.44**	**1.72**	**2.54**	0	0	**1.96**	**36.60**	**10.32**	**5.89**	**7.86**	**0.16**

Table 7.2: **Single-view 3D occupancy prediction results on SSCBench-nuScenes.** We report results of S4C [HWM+24] and our method trained solely on SSCBench-nuScenes ("-nuscs"). We also initialise both methods with SSCBench-KITTI-360 trained models, and report both cross-dataset testing of SSCBench-KITTI-360 trained models ("-kt360") and results of fine-tuned models on SSCBench-nuScenes ("-ft"). Our method achieves state-of-art performance against the existing self-supervised single-view method by a large margin on both settings. (%) indicates each class label ratio. Best numbers in **bold**.

Implementation details. We implement our method in PyTorch and train on two Tesla A40 GPUs. For pseudo semantic labels and fusion, we follow [HWM+24], using ResNet101 Panoptic-Deeplab [CCZ+20] trained on Cityscapes [COR+16] as our frozen semantic segmenter θ_s. Our geometry branch θ'_d and depth estimator θ_d use the ViT-B version of DepthAnythingv2 [YKH+24b]. We train with Adam [KB15] optimiser using a learning rate of 10^{-4} for decoders and 10^{-6} for θ'_d to preserve pretraining. On SSCBench-KITTI-360, we train for 60 epochs, reducing the learning rate 10x after 120k iterations [HWM+24]. On SSCBench-nuScenes, we train for 20 epochs, reducing the learning rate 10x after 20k iterations. For cross-dataset fine-tuning, we initialise with SSCBench-KITTI-360 models and fine-tune on SSCBench-nuScenes for 10 epochs.

7.4.2 3D Occupancy Prediction Results

We show quantitative 3D occupancy prediction results of our method compared with state-of-the-art single-view methods on SSCBench-KITTI-360 in Tab. 7.1 and SSCBench-nuScenes in Tab. 7.2. All the methods feature single-view RGB input, with [CDC22, LYC+23, HZZ+23, ZZD23, HZZ+24b, HTZ+25] trained with groundtruth 3D supervision while [HWM+24] and MT-Occ with 2D supervision. Our method shows improved performance on both datasets, for both geometry reconstruction and semantic segmentation, especially on challenging small objects such as "motocyle", "bicycle", "person".

In addition, SSCBench-nuScenes is significantly smaller and presents a more challenging multi-view training setup, as its front camera has a narrower FoV and lacks front-view stereo pairs,

	Method	IoU (%)	mIoU (%)	car (2.85%)	bicycle (0.01%)	motocycle (0.01%)	truck (0.16%)	other-veh. (5.75%)	person (0.02%)	road (14.98%)	sidewalk (6.43%)	building (15.67%)	fence (0.96%)	vegetation (41.99%)	terrain (7.10%)	pole (0.22%)	traf.-sign (0.06%)	other-obj. (0.28%)
3d-gt-sup	[CDC22]	37.87	13.52	19.34	0.43	0.58	8.02	2.03	0.86	48.35	28.13	32.89	3.53	26.15	16.75	6.92	5.67	3.09
	[LYC+23]	38.76	13.20	17.84	1.16	0.89	4.56	2.06	1.63	47.01	27.21	31.18	4.97	28.99	14.69	6.51	6.92	2.43
	[HZZ+23]	40.22	14.95	21.56	1.09	1.37	8.06	2.57	2.38	52.99	31.07	34.83	4.80	30.08	17.51	7.46	5.86	2.70
	[ZZD23]	**40.27**	14.97	**22.58**	0.66	0.26	9.89	3.82	2.77	**54.30**	31.53	**36.42**	4.80	**31.00**	**19.51**	**7.77**	**8.51**	**4.60**
	[HZZ+24b]	35.38	13.79	18.93	1.02	**4.62**	**18.07**	**7.59**	**3.35**	45.47	25.03	28.44	**5.68**	29.54	8.62	2.99	2.32	5.14
	[HTZ+25]	38.37	**15.33**	21.08	**2.55**	4.21	12.41	5.73	1.59	54.12	**32.31**	32.01	4.98	28.94	17.33	3.57	5.48	3.54
2d-psu-sup	[HWM+24]-kt360	38.84	10.10	10.32	0	0	2.17	0.15	**0.36**	48.61	26.43	20.96	2.80	22.32	16.45	0.43	**0.49**	0
	MT-Occ-kt360	**40.44**	**11.70**	12.75	**2.48**	4.17	**6.90**	**2.74**	0.34	**52.26**	**26.80**	**22.84**	**3.57**	**23.15**	**16.53**	**0.49**	0.34	**0.08**
	[HWM+24]-nuscs	15.11	2.31	2.47	0	0	0.05	0	0.22	8.92	10.43	5.14	0.51	5.98	0.62	0.13	0.15	0.01
	MT-Occ-nuscs	23.77	3.14	2.85	0.24	0.04	0.63	0.93	0.21	8.50	6.85	9.02	1.19	14.21	2.15	0.08	0.19	0
	[HWM+24]-ft-30	34.88	7.55	5.52	0	0	0	0	0	38.37	19.74	17.44	0.26	19.02	12.84	0	0	0
	MT-Occ-ft-30	39.33	10.28	7.02	1.76	1.19	1.79	1.38	0.31	50.46	25.09	21.21	3.20	22.35	17.66	0.33	0.45	0.05
	[HWM+24]-ft-60	38.64	9.99	10.26	0	0	2.17	0.14	0.34	48.63	26.39	20.95	2.82	22.39	16.46	0.45	0.38	0
	MT-Occ-ft-60	40.41	11.68	**12.76**	**2.48**	**4.20**	6.87	2.73	0.34	52.12	26.77	22.80	**3.57**	23.13	16.48	0.48	0.34	**0.08**

Table 7.3: **Cross-dataset 3D occupancy prediction results on SSCBench-KITTI-360.** We report models trained from scratch on SSCBench-KITTI-360 (-kt360), and also initialise both methods with SSCBench-NuScenes trained models, and report both cross-dataset testing of SSCBench-NuScenes trained models ("-nuscs") and results of fine-tuned models on SSCBench-KITTI-360 ("-ft") for 30 epochs ("-30") and 60 epochs ("-60", same number of epochs as from scratch training on SSCBench-KITTI-360). Our method excels in both cross-dataset testing and fine-tuning settings. Best numbers in **bold**.

making image synthesis loss computation more difficult. This is reflected in the lower overall accuracy of models trained solely on this dataset. To evaluate whether large-scale cross-domain data improves accuracy, we initialise both the competing method [HWM+24] and our MT-Occ with models pre-trained on SSCBench-KITTI-360 and report their cross-dataset performance as well as fine-tuned results in Tab. 7.2. Results show that cross-domain models achieve better semantic understanding (mIoU) due to increased training data but struggle with geometry reconstruction (IoU) due to domain gaps, such as differing camera parameters. Fine-tuning improves both metrics. Notably, our method outperforms the competing approach in both cross-dataset testing and fine-tuning, demonstrating superior generalizability.

We further present cross-dataset results where models trained on the larger SSCBench-KITTI-360 dataset are tested and fine-tuned on the smaller SSCBench-NuScenes dataset. To further highlight our method's generalisability, we conduct reverse cross-dataset experiments, initialising models trained on SSCBench-NuScenes and testing and fine-tuning them on SSCBench-KITTI-360. The results are shown in Table 7.3. Our method significantly outperforms S4C [HWM+24] in both cross-dataset testing and fine-tuning scenarios. However, due to the size and diversity difference between the datasets (with SSCBench-NuScenes being much smaller), pretraining on SSCBench-NuScenes does not improve the final accuracy on SSCBench-KITTI-360. This could be due to the "forgetting" issue commonly observed in source-free domain adaptation tasks [LSSD23]. Note that designing effective domain adaptation techniques in source-free,

	Method	IoU (%)	mIoU (%)	car (2.85%)	bicycle (0.01%)	motocycle (0.01%)	truck (0.16%)	other-veh. (5.75%)	person (0.02%)	road (14.98%)	sidewalk (6.43%)	building (15.67%)	fence (0.96%)	vegetation (41.99%)	terrian (7.10%)	pole (0.22%)	traf.-sign (0.06%)	other-obj. (0.28%)
visible	[HWM+24]	23.04	9.68	7.53	0	0	3.09	0.25	**0.44**	54.95	25.25	18.11	3.89	11.71	**18.37**	0.69	**0.96**	0
	MT-Occ	**26.64**	**11.15**	**9.08**	**3.05**	**4.27**	**5.69**	**1.60**	0.43	**57.28**	**29.59**	**20.29**	**4.43**	**13.98**	15.48	**0.98**	0.93	**0.22**
occl.	[HWM+24]	42.31	10.19	11.08	0	0	2.05	0.12	**0.34**	47.08	**26.72**	21.46	2.60	24.47	16.25	0.31	**0.31**	0
	MT-Occ	**43.18**	**11.78**	**13.64**	**2.25**	**4.13**	**7.21**	**3.10**	0.30	**50.92**	26.36	**23.24**	**3.42**	**24.78**	**16.70**	**0.32**	0.20	**0.06**

Table 7.4: **Visible and invisible area breakdown results on SSCBench-KITTI-360.** Our method improves on both visible surfaces and occluded areas. (%) indicates each class label ratio. Best numbers in **bold**.

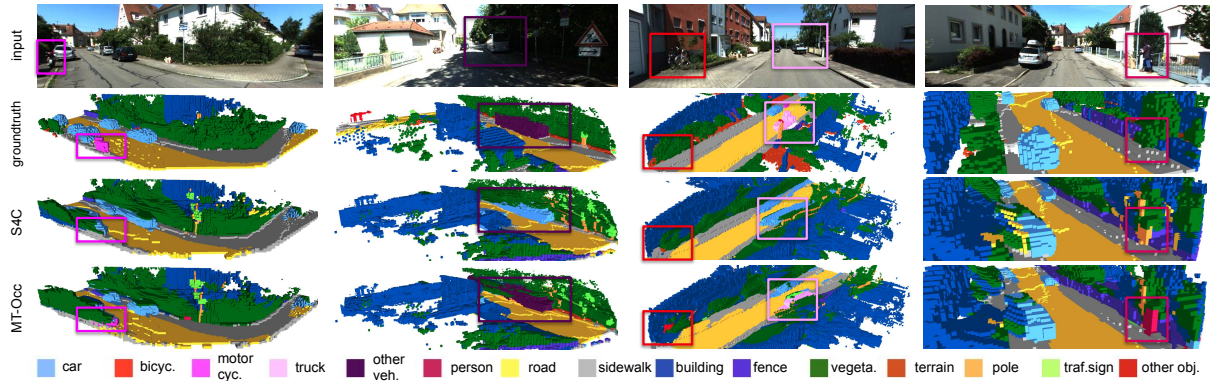


Figure 7.4: **Qualitative results on SSCBench-KITTI-360.** Compared to S4C[HWM+24], our MT-Occ captures more accurate geometry and semantics, especially on challenging small classes or rare classes.

cross-dataset fine-tuning setups is challenging and presents an interesting direction for future research.

We divide the scenes into visible surface and invisible areas by applying ray tracing techniques [PJH23] on the SSCBench-KITTI-360 3D occupancy groundtruth to generate visibility masks, and calculate the 3D occupancy prediction performance respectively for visible and occluded parts, shown in Tab. 7.4. Our method consistently improves on both visible and occluded areas in the scene, and yields even more improvements in the occluded areas of vehicle classes such as "car", "truck" and "other vehicles", indicating more accurate geometry reconstruction for safety critical classes.

We present qualitative results in Fig. 7.4. Our method achieves more accurate scene geometry reconstruction, particularly in unseen areas where existing methods often produce "trailing artifacts" – objects (e.g., cars) appearing stretched along the camera's viewing direction in self-occluded regions. This improvement is due to our spatial cross-attention design. Additionally, our approach better captures small or rare objects.

7.4.3 2D Semantic Rendering Results

We further evaluate the quality of 2D rendered semantics against ground-truth 2D semantic maps of front views from SSCBench-KITTI-360. In Tab. 7.5, we compare the quality of the semantic pseudo-labels used for training and the rendered semantics of each method. Our approach

Method	mIoU (%)	car (2.85%)	bicycle (0.01%)	motocycle (0.01%)	truck (0.16%)	other-veh. (5.75%)	person (0.02%)	road (14.98%)	sidewalk (6.43%)	building (15.67%)	fence (0.96%)	vegetation (41.99%)	terrian (7.10%)	pole (0.22%)	traf.-sign (0.06%)	other-obj. (0.28%)
[CCZ+20]	*46.55*	*86.49*	*21.57*	*35.09*	*36.95*	*10.94*	*28.38*	*83.55*	*57.55*	*83.47*	*44.00*	*85.02*	*55.22*	*32.75*	*37.13*	*0.18*
[HWM+24]	39.04	82.72	0	0	23.78	0.34	9.91	**87.01**	60.31	**84.38**	41.13	**85.03**	**60.60**	24.50	25.97	0
MT-Occ	**46.50**	**84.87**	**21.58**	**35.17**	**36.80**	**11.20**	**26.60**	86.42	**62.42**	83.97	**44.21**	84.98	56.77	**29.47**	**32.84**	**0.19**

Table 7.5: **Rendered 2D semantic segmentation results on SSCBench-KITTI-360.** Our method outperforms the existing method by a large margin especially on challenging small classes, and yields even comparative results to the 2D pseudo labels both methods use to train (PanopticDeeplab[CCZ+20]). (%) indicates each class label ratio. Best numbers in **bold**.

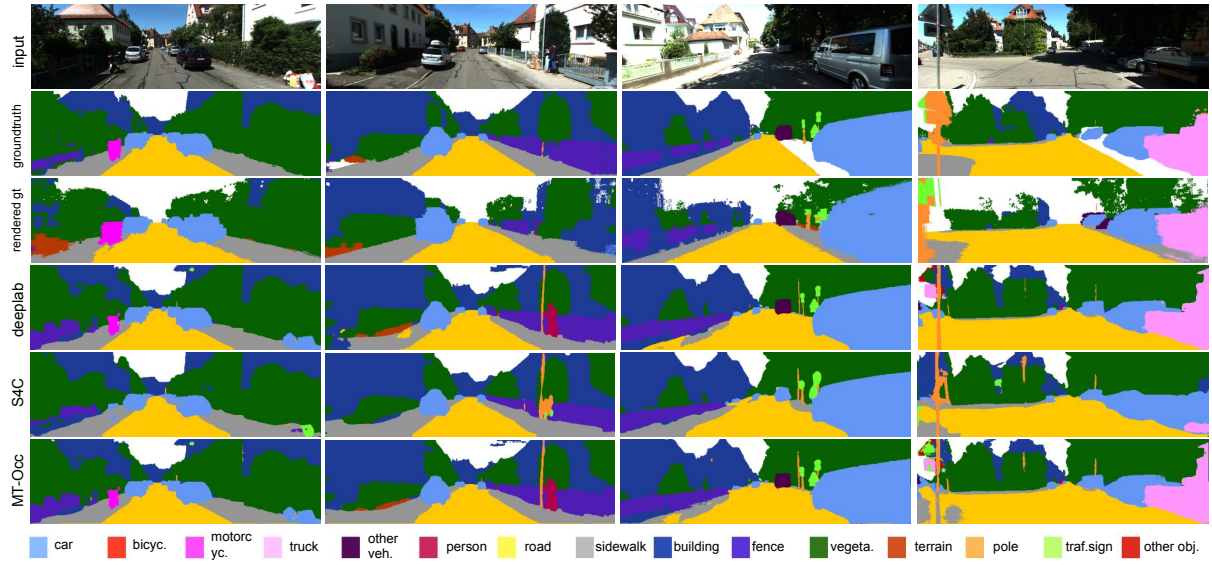


Figure 7.5: **Rendered semantic segmentation results on SSCBench-KITTI-360.** MT-Occ produces more faithful renderings to pseudo groundtruth (deeplab) and manually annotated 2D semantic groundturth (groundtruth).

significantly outperforms the existing method, achieving results close to the pseudo ground truth, especially for challenging small classes like "motorcycle", "bicycle", and "person". Notably, due to SSCBench-KITTI-360's sparse LiDAR aggregation, some small-class voxels (e.g., "person") may be missing in the 3D ground truth. However, 2D rendering results show that our method greatly improves the capture of these classes. The substantial improvement in 2D rendered semantics suggests that our method more effectively distils 2D information.

Furthermore, we show some qualitative results of the rendered semantic segmentation of MT-Occ compared to S4C [HWM+24] in Fig. 7.5. We can see that MT-Occ more faithfully renders plausible 2D segmentation maps, and yields closer results to both the 2D semantic segmentation groundtruth and the pseudo semantic labels from Panoptic-Deeplab [CCZ+20] used for both methods, especially on challenging classes. Thanks to the feature fusion strategy, MT-Occ renderings are very close to the Panoptic-Deeplab pseudo groundtruth, indicating that our method is able to distil more information from 2D.

feature fusion		semantic refine			attn	**IoU**	**mIoU**
θ'_d	θ_s	$\mathcal{L}_{\text{depth}}$	S'	u'			
S4C [HWM+24]						38.84	10.10
dino_v2 [ODM+25]						38.20	9.77
da_v2 [YKH+24b]						39.99	10.66
da_v2 [YKH+24b]	✓					40.09	11.03
res50 [HWM+24]		✓	✓	✓		40.06	10.58
da_v2 [YKH+24b]		✓				39.89	10.73
da_v2 [YKH+24b]		✓	✓			40.03	10.90
da_v2 [YKH+24b]		✓	✓	✓		40.15	11.06
da_v2 [YKH+24b]	✓	✓	✓	✓		40.30	11.34
res50 [HWM+24]					✓	40.15	10.65
res50 [HWM+24]		✓	✓	✓	✓	40.14	11.07
da_v2 [YKH+24b]					✓	40.55	11.07
da_v2 [YKH+24b]	✓				✓	**40.61**	11.26
da_v2 [YKH+24b]		✓	✓	✓	✓	40.50	11.30
da_v2 [YKH+24b]	✓	✓	✓	✓	✓	40.44	**11.72**

Table 7.6: **Ablation results on SSCBench-KITTI-360.** θ'_d: pretrained encoder; θ_s: semantic fusion; $\mathcal{L}_{\text{depth}}$: depth guidance loss; S': pseudo label refinement; u': semantic-guided sampling; "attn": spatial cross-task attention. Best numbers in **bold**.

7.4.4 Ablation Study

We conduct an ablation study on our proposed components for feature fusion, semantic refinement, and spatial cross-task attention in Tab. 7.6. The results confirm that each component contributes to performance gains, with the full model achieving the best results. The robust performance gain of each combination of the components also indicates that the proposed components are also flexible to work stand alone, as plug-and-play to similar architectures or frameworks for single-view 3D occupancy prediction.

Additionally, we compare the use of the DepthAnythingv2 encoder against the DINOv2 encoder. As shown in Tab. 7.6, DepthAnythingv2 provides significant improvements, whereas DINOv2 does not. This highlights the importance of depth pretraining for this task, rather than just architectural changes.

7.5 Conclusion

In this paper, we address the challenging task of jointly reconstructing scene geometry and semantics from a single view and introduce MT-Occ, a novel approach for single-view 3D occupancy prediction with effective multi-task distillation. By incorporating components for multi-task feature fusion, spatial cross-task attention, and view-consistent pseudo-label refinement, MT-Occ significantly outperforms existing methods. It excels in capturing challenging small

objects that are underrepresented in the training set and achieves more realistic, semantic-aware geometry reconstruction, particularly in self-occluded regions. Additionally, MT-Occ demonstrates enhanced cross-dataset generalisation, paving the way for a versatile 3D occupancy prediction framework that can be learned from large-scale unlabelled 2D data.

While MT-Occ provides a valuable approach for developing annotation-free, low-cost systems, it still has limitations that suggest promising directions for future research. Though improved by our proposed components, the accuracy of 2D-supervised methods remains lower than that of methods relying on 3D occupancy annotations. While avoiding the need for 3D ground truth is a significant advantage, further exploration of how large-scale unannotated data, such as aggregating datasets from different domains, could further improve accuracy is worth investigating.

Second, although the proposed components allow for flexibility in trading off performance for inference speed, being in line with NeRF-based frameworks [MST+21], our method does not yet support real-time inference. Future research could explore runtime optimisation techniques or more efficient scene representation methods, such as 3DGS [KKLD23], as potential solutions.

8 Conclusion and Future Work

Contents

Recent progress in 3D vision has brought monocular scene understanding closer to practical deployment across applications such as autonomous driving, robotics, and augmented reality. However, learning holistic 3D representations from monocular inputs remains highly challenging due to limited geometric cues, lack of dense supervision, and domain shifts at test time. These constraints are particularly pronounced in real-world scenarios, where expensive depth sensors, precise labels, or controlled capture environments are rarely available. This motivates a shift in methodology from supervision-heavy pipelines toward more scalable solutions that combine model-based priors, multi-task learning, and self-supervised adaptation. Moreover, practical deployment demands generalisation across diverse and unseen environments, as well as robustness to noisy, sparse, or ambiguous inputs. To this end, this thesis proposes monocular methods that explicitly reason about geometry, semantics, and contact in dynamic human-scene interactions, while leveraging test-time adaptation and novel scene representations to reduce supervision requirements and increase scalability in the wild.

8.1 Key Insights and Conclusions

We addressed the challenge of robust 3D perception from monocular images through two complementary research directions. In **Part I**, we focused on the problem of human motion understanding in realistic scenes. We investigated monocular 3D human pose and motion estimation, moving beyond isolated predictions to scene-aware motion capture. In particular, we studied how physical plausibility, body-scene contact, and environmental constraints could be jointly leveraged to reduce depth ambiguity and improve global pose estimation in complex environments. In **Part II**, we turned to scene reconstruction from monocular RGB input, framing depth estimation as a key modality for understanding 3D structure. Recognising that depth is not only a low-level cue but also a form of geometric reconstruction, we explored test-time adaptation strategies that enable robust generalisation to unseen domains without requiring source data. Building further upon this, we studied semantic scene completion and 3D occupancy prediction, and proposed self-supervised learning frameworks based on neural rendering and Gaussian representations. These approaches enabled us to infer high-resolution voxel-level semantics and geometry from vision-only inputs, reducing the reliance on dense supervision or specialised hardware.

In the following, we revisit the contributions of each chapter in more detail, before discussing future directions in Section 8.2.

Multi-person Pose Estimation in Scene Scales. In Chapter 3, we addressed the problem of multi-person 3D pose estimation from monocular images at scene scale, where accurate absolute positioning is critical for downstream applications such as human-robot interaction or surveillance. Existing monocular methods typically focus on relative pose within a local

coordinate system and struggle with global localisation due to the lack of scale cues. We proposed a correction-factor-based framework that disentangles human pose from body scale, enabling improved global localisation across multiple subjects. By leveraging a scale-aware formulation guided by 2D projection statistics, our method captures inter-person relationships and enhances scene-level spatial reasoning. This formulation not only improves absolute position accuracy but also enhances the consistency of multi-person interactions within the 3D scene, all without relying on external depth sensors or multi-view setups.

Human Motion Capture from Scene Contact Guidance. Chapter 4 explored the integration of physical contact cues into monocular 3D human motion capture. While previous approaches often treat the human body as isolated from its environment, this assumption leads to implausible results such as floating or foot-sliding, particularly in cluttered or occluded scenes. We proposed a contact-aware optimisation framework that reasons explicitly about dense surface-level interactions between the human body and the surrounding scene. By sampling diverse candidate motions from a learned pose manifold and refining them via contact-consistent optimisation, the method is able to disambiguate depth and improve absolute-scale estimation. Notably, the incorporation of physical plausibility through body–scene contact significantly enhances the realism and stability of the recovered motion, without requiring ground-truth 3D annotations or depth maps. This chapter marks an important step toward bridging the gap between visual perception and physical scene understanding in monocular settings.

Human Motion Capture with Scene Deformation Recovery. Chapter 5 advanced monocular motion capture by addressing the critical yet underexplored problem of scene deformability. Most existing monocular MoCap systems assume rigid environments and thereby fail to capture the nuanced interactions that occur when the human body comes into contact with soft or movable surfaces—such as sitting on a sofa or kneeling on a mattress. To resolve this, we introduced a joint optimisation framework that estimates both human pose and non-rigid scene deformation from monocular RGB input. Our method identifies dense body–scene contact points via ray-casting, then adjusts the underlying scene geometry to align with physically plausible support. By treating the scene not as a static constraint but as a dynamic, deformable entity, this chapter presents a more holistic model of human–environment interaction. The result is improved localisation accuracy, reduced interpenetration, and enhanced realism—without requiring RGB-D input, 3D supervision, or additional sensing. This work pushes monocular motion capture closer to functioning reliably in real-world, unconstrained environments.

Test-Time Adaptation for Monocular Depth Estimation. Chapter 6 explores the challenge of deploying monocular depth estimation models in unseen domains, where shifts in lighting, weather, or sensor characteristics can drastically reduce model performance. While supervised or self-supervised monocular depth estimation has made significant progress, most existing methods rely on static training–testing boundaries and cannot generalise robustly across domains. To address this, we proposed a test-time adaptation (TTA) framework that allows a pretrained depth model to self-adapt during inference, without requiring access to source data or any ground-truth labels. Our method leverages geometric and photometric consistency as self-supervision signals, updating the model online to better align with the test-time domain. Crucially, we achieved this using lightweight architectural modifications and minimal computational overhead, making the method practical for real-world deployment scenarios such as autonomous driving. By treating depth estimation as a form of scene reconstruction, this chapter demonstrates that robust, domain-adaptive 3D understanding is achievable from monocular input—even under challenging, dynamic conditions.

Single-View Semantic Occupancy Prediction with Multi-Task Supervision. Chapter 7 addresses the challenging problem of reconstructing semantically meaningful 3D occupancy

grids from a single RGB image. Unlike methods that require LiDAR, multi-view input, or dense 3D ground truth, our approach relies solely on monocular imagery and 2D supervision, making it highly practical for settings with minimal sensor setups. To this end, we introduced a differentiable volumetric rendering framework inspired by neural implicit representations, allowing 3D voxel predictions to be trained using 2D semantic maps through rendering-based losses. Moreover, we proposed a multi-task interaction module that jointly learns depth and semantic segmentation as auxiliary tasks, encouraging rich 2D-to-3D feature transfer. This joint learning strategy significantly improves the geometric consistency and semantic accuracy of the predicted 3D occupancy, even in the absence of any 3D supervision. By demonstrating that accurate 3D scene understanding can emerge from single-frame monocular inputs, this chapter reinforces the broader thesis goal of enabling lightweight and scalable 3D perception under minimal assumptions.

Together, the chapters of this thesis demonstrate how structured representations, scene-aware constraints, and multi-task supervisory signals can significantly enhance 3D perception from monocular inputs. By progressively moving from human-centric reasoning to full-scene reconstruction, our contributions bridge the gap between local pose estimation and holistic scene understanding. Across both parts, we show that carefully designed priors—whether geometric, semantic, or physical—can compensate for the inherent ambiguity of monocular data, enabling globally consistent and physically plausible 3D predictions. Beyond technical accuracy, these works reflect a broader shift towards more scalable, annotation-efficient, and generalisable 3D perception systems, laying the groundwork for future research in resource-constrained, real-world deployment settings.

8.2 Future Directions

This thesis explored monocular 3D human understanding and scene reconstruction through the lens of learning-based representations, contact reasoning, and domain adaptation. While each contribution addresses specific technical challenges, the broader landscape is evolving rapidly with new developments in scalable architectures, unified representations, and pre-trained foundation models. Below, we outline several future directions aligned with the two thematic parts of the thesis.

Generalisable and Scalable Human Modelling with Vision Foundations. While existing approaches typically infer human-scene contact and interaction from geometric evidence alone, future systems stand to benefit significantly from the integration of high-level cues such as semantics, affordances, and physical plausibility. Foundation models like SAM [KMR+23] and DepthAnything [YKH+24a, YKH+24b] offer rich intermediate representations—e.g., segmentation masks, depth priors—that can be incorporated as supervisory signals or structural constraints to enhance contact reasoning. Moreover, scaling to multi-human scenarios introduces further challenges such as mutual occlusion and entangled interactions. Here, pretrained vision backbones such as DINO may support dense part-level matching or correspondence reasoning, helping to disentangle individual trajectories and interactions from monocular input. Taken together, these models can provide a strong basis for generalisable, data-efficient, and socially-aware motion capture frameworks.

From Individual Tasks to Unified Human Modelling. While this thesis has treated pose estimation, motion capture, and contact reasoning as related yet distinct problems, a compelling future direction lies in unifying these tasks within a holistic spatio-temporal framework. Such a formulation would jointly model static attributes (e.g., shape and identity) alongside dynamic behaviours (e.g., motion, contact events), enabling greater consistency over time and improved

robustness to occlusions or ambiguities. Incorporating deformation modelling into this unified setup would further enhance realism, particularly in non-rigid scenes or when the environment itself is subject to change. For example, learning to capture subtle scene deformations caused by human motion—such as soft furniture or elastic surfaces—could lead to more physically plausible reconstructions. Moreover, recent advances in neural implicit representations and temporally consistent priors provide promising tools for developing such integrated systems. In the long term, these efforts may support richer human-scene understanding for downstream tasks like simulation, behavioural analysis, or interactive robotics.

Domain Adaptation in the Era of Foundation Models. The emergence of vision foundation models—such as DINO [ODM+25] for representation learning, SAM [KMR+23] for segmentation, and DepthAnything [YKH+24a, YKH+24b] for dense prediction—has shifted the landscape of domain adaptation. Rather than adapting a task-specific model to each deployment scenario, future frameworks may instead harness the generalisation capabilities of pretrained models as a starting point. This opens several avenues: one is to use foundation models as frozen feature extractors, building lightweight adaptation layers on top to account for domain-specific nuances. Another is to perform continual adaptation at the feature level via modulation or adapter tuning, thus avoiding full model retraining. Importantly, in tasks such as monocular depth estimation, pretrained depth priors (e.g., from DepthAnything [EPF14, YKH+24b] or similar models) may serve as universal anchors, enabling fast and stable test-time adaptation even without source data. More broadly, the focus may shift from raw domain alignment to task transfer and semantic consistency, capitalising on the structured outputs and strong inductive biases embedded within large-scale pretraining.

Lifting 2D Perception to 3D Scene Reconstruction. This thesis highlights that 2D perception tasks—such as depth estimation, semantic segmentation, surface normal prediction, and motion analysis—can serve not only as standalone objectives but also as essential intermediate cues for reconstructing complete 3D scenes. Monocular inputs alone are inherently ambiguous, making 3D reasoning ill-posed; however, structured 2D supervision can provide rich spatial and semantic constraints. A key direction for future research is to unify these modalities within multi-task learning frameworks that regularise 3D predictions, particularly in self-supervised or weakly supervised settings. Rather than learning from scratch, future systems may benefit from distilling knowledge from specialised 2D models or leveraging large pre-trained foundation models. Notably, geometry-aware 3D backbones like VGGT [WCK+25] offer the potential to encode such priors directly in 3D space, improving generalisability and efficiency in sparse-label or real-world scenarios.

Accelerating 3D Reconstruction Beyond Neural Rendering. While neural implicit fields like NeRF [MST+21] have demonstrated impressive fidelity in 3D reconstruction, their reliance on dense volumetric sampling and MLP-based querying severely limits inference speed—posing challenges for real-time applications. Recent advances such as 3D Gaussian Splatting (3DGS) [KKLD23] offer a compelling alternative by replacing slow ray marching with efficient rasterisation of Gaussian primitives, achieving real-time rendering with significantly reduced compute. However, 3DGS may trade off some geometric precision compared to NeRF. A promising direction lies in exploring hybrid architectures that combine the speed of splatting with the accuracy of neural fields, potentially guided by geometry-aware priors from 3D foundation models like VGGT [WCK+25]. Such designs could unlock fast, scalable 3D perception for practical deployment scenarios.

Across the directions outlined above, a common theme emerges: the move from task-specific, label-intensive systems towards unified, data-efficient, and generalisable frameworks. Whether for human-centric perception or scene-level understanding, future research is likely to benefit

from the interplay between modular task designs, powerful pretrained backbones, and structured multi-modal supervision. Foundation models act not only as performance boosters, but also as enablers of new paradigms—facilitating rapid adaptation, scaling to diverse environments, and bridging the gap between low-level signals and high-level understanding. Building on the contributions of this thesis, future work has the opportunity to further blur the lines between geometry, semantics, and dynamics, pushing towards more holistic and deployable 3D perception systems.

List of Figures

List of Tables

Bibliography

[AB15] Ijaz Akhter and Michael J Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. Cited on page 2.

[ADZ19] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. Cited on page 12.

[AGR+16] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. Cited on page 2.

[AVIL21] Shubhra Aich, Jean Marie Uwabeza Vianney, Md Amirul Islam, and Mannat Kaur Bingbing Liu. Bidirectional attention network for monocular depth estimation. In *Proceedings of the IEEE International Conference on Robotics and Animation (ICRA)*. IEEE, 2021. Cited on pages 14 and 70.

[AWS+22] Hiroyasu Akada, Jian Wang, Soshi Shimada, Masaki Takahashi, Christian Theobalt, and Vladislav Golyanik. UnrealEgo: A new dataset for robust egocentric 3d human motion capture. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. Cited on page 51.

[BAW21] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. Cited on page 72.

[BBLR15] Federica Bogo, Michael J. Black, Matthew Loper, and Javier Romero. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. Cited on page 11.

[BCW+20] Yanrui Bin, Zhao-Min Chen, Xiu-Shen Wei, Xinya Chen, Changxin Gao, and Nong Sang. Structure-aware human pose estimation with graph convolutional networks. *Pattern Recognition*, 106:107410, 2020. Cited on page 11.

[BGR25] Simon Boeder, Fabian Gigengack, and Benjamin Risse. Gaussianflowocc: Sparse and weakly supervised occupancy estimation using gaussian splatting and temporal flow. *arXiv preprint arXiv:2502.17288*, 2025. Cited on pages 16 and 86.

[BKK22] Jongbeom Baek, Gyeongnyeon Kim, and Seungryong Kim. Semi-supervised learning with mutual distillation for monocular depth estimation. In *Proceedings of the IEEE International Conference on Robotics and Animation (ICRA)*, 2022. Cited on page 71.

[BKL+16] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. Cited on pages 2, 3, 11, 37, 39, 53, and 55.

[BLW+19] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. Cited on page 6.

[BMAZ22] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022. Cited on pages 17 and 86.

[BTHD18] Andreea Bobu, Eric Tzeng, Judy Hoffman, and Trevor Darrell. Adapting to continuously shifting domains. In *Proceedings of the International Conference on Learning Representations Workshops (ICLRW)*, 2018. Cited on pages 15 and 70.

[BTT+20] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020. Cited on page 13.

[BVGH19] Lewis Bridgeman, Marco Volino, Jean-Yves Guillemaut, and Adrian Hilton. Multi-person 3d pose estimation and tracking in sports. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. Cited on page 22.

[BZSS22] Deblina Bhattacharjee, Tong Zhang, Sabine Süsstrunk, and Mathieu Salzmann. Mult: An end-to-end multitask learning transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. Cited on pages 17 and 86.

[CAR+21] Ran Cheng, Christopher Agia, Yuan Ren, Xinhai Li, and Liu Bingbing. S3cnet: A sparse semantic scene completion network for lidar point clouds. In *Proceedings of the Annual Conference on Robot Learning (CoRL)*. PMLR, 2021. Cited on pages 16 and 85.

[CBL+20] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Cited on page 91.

[CCZ+20] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Cited on pages 92, 95, and 110.

[CDC22] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. Cited on pages 16, 85, 86, 88, 91, 92, and 93.

[CDdC24] Anh-Quan Cao, Angela Dai, and Raoul de Charette. Pasco: Urban 3d panoptic scene completion with uncertainty awareness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. Cited on page 85.

[CGM+20] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. Cited on pages 13, 39, 40, 45, 54, and 105.

[CHL+24] Zi-Ting Chou, Sheng-Yu Huang, I Liu, Yu-Chiang Frank Wang, et al. Gsnerf: Generalizable semantic neural radiance fields with enhanced 3d scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. Cited on pages 16 and 86.

[CHS+19] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 43(1):172–186, 2019. Cited on pages 12, 22, 23, and 43.

[CL18] Ju Yong Chang and Kyoung Mu Lee. 2d–3d pose consistency-based conditional random fields for 3d human pose estimation. *Computer Vision and Image Understanding (CVIU)*, 169:52–61, 2018. Cited on pages 23 and 24.

[CML21] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. Cited on pages 12, 37, and 39.

[COR+16] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. Cited on page 92.

[CPB+20] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020. Cited on page 12.

[CPEZ13] James Charles, Tomas Pfister, Mark Everingham, and Andrew Zisserman. Automatic and efficient human pose estimation for sign language videos. *International Journal of Computer Vision (IJCV)*, 110:70–90, 10 2013. Cited on page 40.

[CR17] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. Cited on pages 11, 13, 23, 24, 28, 29, 37, 39, and 53.

[CSS19] Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. Cited on pages 15 and 70.

[CSWS17] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. Cited on page 55.

[CZB+25] Loick Chambon, Eloi Zablocki, Alexandre Boulch, Mickael Chen, and Matthieu Cord. Gaussrender: Learning 3d occupancy with gaussian rendering. *arXiv preprint arXiv:2502.05040*, 2025. Cited on pages 16 and 86.

[DCS+17] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. Cited on pages 5 and 58.

[DGM+19] Rishabh Dabral, Nitesh B Gundavarapu, Rahul Mitra, Abhishek Sharma, Ganesh Ramakrishnan, and Arjun Jain. Multi-person 3d human pose estimation from monocular images. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2019. Cited on pages 12, 13, 51, and 53.

[DJH+19] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. Cited on page 22.

[DMK+18] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. Cited on page 12.

[DSJ+21] Rishabh Dabral, Soshi Shimada, Arjun Jain, Christian Theobalt, and Vladislav Golyanik. Gravity-aware monocular 3d human-object reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. Cited on pages 43, 51, and 53.

[ECC+21] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. Cited on pages 76, 78, 80, and 110.

[EPF14] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems (NeurIPS)*, 27, 2014. Cited on pages 1, 14, 70, 72, 76, 77, 78, 90, 102, and 110.

[FAZ+21] Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021. Cited on page 86.

[FXTL17] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. Cited on pages 12 and 22.

[FXW+18] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *Proceedings of the conference on artificial intelligence (AAAI)*, volume 32, 2018. Cited on pages 13, 22, 23, 24, 29, and 53.

[FYJ+22] Ziyue Feng, Liang Yang, Longlong Jing, Haiyan Wang, YingLi Tian, and Bing Li. Disentangling object motion and occlusion for unsupervised multi-frame monocular depth. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022. Cited on pages 16 and 86.

[FZC+22] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, 2022. Cited on pages 16 and 86.

[FZO+20] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Cited on page 39.

[GAP+20] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Cited on pages 15, 70, 76, 77, and 110.

[Gav99] Dariu M Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding (CVIU)*, 73(1):82–98, 1999. Cited on page 53.

[GEM87] Stuart Geman and Donald E. McClure. Statistical methods for tomographic image reconstruction. In *46th Session of the International Statistical Institute (ISI)*, volume 4, pages 5–21, 1987. Cited on pages 56, 57, and 59.

[Gib50] James J Gibson. *The perception of the visual world*. Houghton Mifflin, 1950. Cited on page 51.

[GL15] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2015. Cited on pages 15 and 69.

[GLK+20] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020. Cited on page 12.

[GLSU13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 32(11):1231–1237, 2013. Cited on pages 6 and 76.

[GLX+24] Wanshui Gan, Fang Liu, Hongbin Xu, Ningkai Mo, and Naoto Yokoya. Gaussianocc: Fully self-supervised and efficient 3d occupancy estimation with gaussian splatting. *arXiv preprint arXiv:2408.11447*, 2024. Cited on pages 16 and 86.

[GMAB17] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. Cited on page 73.

[GMAFB19] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. Cited on pages 6, 15, 70, 72, 73, 77, 79, and 90.

[GML+21] Yu Guo, Lichen Ma, Zhi Li, Xuan Wang, and Fei Wang. Monocular 3d multi-person pose estimation via predicting factorized correction factors. *Computer Vision and Image Understanding (CVIU)*, 213:103278, 2021. Cited on pages 8, 9, and 21.

[GMSPM21] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. Cited on page 52.

[HBS+21] Zeyu Hu, Xuyang Bai, Jiaxiang Shang, Runze Zhang, Jiayu Dong, Xin Wang, Guangyuan Sun, Hongbo Fu, and Chiew-Lan Tai. Vmnet: Voxel-mesh network for geodesic-aware 3d semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. Cited on pages 58 and 60.

[HCTB19] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. Cited on pages 3, 13, 14, 38, 39, 45, 46, 47, 49, 50, 51, 52, 53, 54, 55, 56, 60, 61, 62, 105, 106, and 109.

[HCV+21] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. Cited on page 39.

[HDVG22] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. Cited on pages 15 and 69.

[HGDG17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. Cited on pages 27, 31, and 33.

[HGT17] Shaoli Huang, Mingming Gong, and Dacheng Tao. A coarse-fine network for keypoint localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. Cited on page 22.

[HGT+21a] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. Cited on pages 13, 39, 45, 46, 47, and 50.

[HGT+21b] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. Cited on pages 13, 53, 55, 56, 57, 61, and 62.

[HL18] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. Cited on page 12.

[HLX+23] Marc Habermann, Lingjie Liu, Weipeng Xu, Gerard Pons-Moll, Michael Zollhoefer, and Christian Theobalt. Hdhumans: A hybrid approach for high-fidelity digital humans. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(3):1–23, 2023. Cited on page 12.

[HTP+18] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*. Pmlr, 2018. Cited on pages 15 and 69.

[HTZ+25] Yuanhui Huang, Amonnut Thammatadatrakoon, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, and Jiwen Lu. Probabilistic gaussian superposition for efficient 3d occupancy prediction. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. Cited on pages 16, 85, 91, 92, and 93.

[HUC+21] Xuefeng Hu, Gokhan Uzunbas, Sirius Chen, Rui Wang, Ashish Shah, Ram Nevatia, and Ser-Nam Lim. Mixnorm: Test-time adaptation through online normalization estimation. *arXiv preprint arXiv:2110.11478*, 2021. Cited on pages 15 and 69.

[HVT+19] Yana Hasson, Gül Varol, Dimitrios Tzionas, Federica Bogo, Ivan Laptev, Cordelia Schmid, and Michael J Black. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. Cited on pages 2, 3, 4, and 12.

[HWM+24] Adrian Hayler, Felix Wimbauer, Dominik Muhle, Christian Rupprecht, and Daniel Cremers. S4c: Self-supervised semantic scene completion with neural fields. *Proceedings of the International Conference on 3D Vision (3DV)*, 2024. Cited on pages 5, 16, 84, 85, 86, 88, 89, 90, 91, 92, 93, 94, 95, 96, 107, and 110.

[HXM+19] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. Cited on pages 13, 37, 39, and 53.

[HXZ+19] Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions On Graphics (TOG)*, 38(2):14:1–14:17, 2019. Cited on page 53.

[HXZ+20] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Cited on page 53.

[HXZ+21] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. A deeper look into deepcap. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2021. Cited on page 53.

[HZ03] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. Cited on pages 4, 6, and 73.

[HZF21] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. Cited on pages 22 and 31.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. Cited on pages 27, 31, 72, and 88.

[HZZ+23] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. Cited on pages 91, 92, and 93.

[HZZ+24a] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. Selfocc: Self-supervised vision-based 3d occupancy prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. Cited on pages 6, 16, 85, 86, and 89.

[HZZ+24b] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2024. Cited on pages 16, 85, 91, 92, and 93.

[IPOS13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(7):1325–1339, 2013. Cited on pages 2 and 27.

[JKP+20] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Cited on page 39.

[JSS18] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. Cited on page 12.

[JTM10] Vijay John, Emanuele Trucco, and Stephen McKenna. Markerless human motion capture using charting and manifold constrained particle swarm optimisation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2010. Cited on page 40.

[JY17] Ehsan Jahangiri and Alan L Yuille. Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017. Cited on pages 22 and 29.

[KAB20] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Cited on pages 12, 13, 37, 39, 53, and 55.

[KB15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. Cited on pages 31, 60, and 92.

[KBJM18] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. Cited on pages 2, 11, 12, 13, 37, 39, 51, 53, and 55.

[KECK21] Neerav Karani, Ertunc Erdil, Krishna Chaitanya, and Ender Konukoglu. Test-time adaptable neural networks for robust medical image segmentation. *Medical Image Analysis*, 68:101907, 2021. Cited on pages 15 and 69.

[KGY+22] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. Cited on pages 16 and 86.

[KHHB21] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. Cited on pages 37 and 39.

[KHT+21] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. Cited on pages 37 and 39.

[KKG+23] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. Cited on pages 16 and 86.

[KKLD23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4):139–1, 2023. Cited on pages 7, 16, 86, 97, and 102.

[KLSW09] Christian Knauer, Maarten Löffler, Marc Scherfenberg, and Thomas Wolle. The directed hausdorff distance between imprecise point sets. In *International Symposium on Algorithms and Computation (ISAAC)*, 2009. Cited on page 43.

[KMR+23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. Cited on pages 85, 101, and 102.

[KPBD19] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. Cited on pages 2, 11, 12, 37, and 39.

[KPJD21] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. Cited on pages 12, 37, and 39.

[KPVG21] Yevhen Kuznietsov, Marc Proesmans, and Luc Van Gool. Comoda: Continuous monocular depth adaptation using past experiences. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021. Cited on pages 5, 15, 68, 69, 70, 77, 78, and 79.

[KSN21] Vinod K Kurmi, Venkatesh K Subramanian, and Vinay P Namboodiri. Domain impression: A source data free domain adaptation method. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021. Cited on pages 15 and 69.

[KTEM18] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. Cited on page 16.

[KVB+20] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Cited on pages 15 and 69.

[KW14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. Cited on page 44.

[KWHG20] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Cited on page 60.

[KZFM19] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. Cited on pages 37 and 39.

[LARS25] Zhi Li, Rahaf Aljundi, Daniel Olmeda Reino, and Bernt Schiele. Mt-occ: single-view 3d occupancy prediction via multi-task distillation. In *Proceedings of the DAGM German Conference on Pattern Recognition*. Springer, 2025. Cited on pages 9, 10, and 83.

[LC14] Sijin Li and Antoni B Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Springer, 2014. Cited on page 23.

[LCWJ15] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2015. Cited on pages 15 and 69.

[LFS+24] Rui Li, Tobias Fischer, Mattia Segu, Marc Pollefeys, Luc Van Gool, and Federico Tombari. Know your neighbors: Improving single-view reconstruction via spatial vision-language reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. Cited on pages 16, 85, 86, and 89.

[LHF20] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2020. Cited on pages 15 and 69.

[LHKS19] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. Cited on pages 6, 14, 70, and 72.

[LHS+20] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics (ToG)*, 39(4):71–1, 2020. Cited on pages 15 and 70.

[LHW+20] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Cited on pages 16 and 85.

[LHZ+18] Shice Liu, Yu Hu, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, and Xiaowei Li. See and think: Disentangling semantic scene completion. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018. Cited on pages 16 and 85.

[LJC+20] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Cited on pages 15 and 69.

[LKP+20] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Cited on pages 28 and 33.

[LLC+21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. Cited on page 71.

[LLDG19] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. Cited on pages 15 and 69.

[LLG+19] Jie Li, Yu Liu, Dong Gong, Qinfeng Shi, Xia Yuan, Chunxia Zhao, and Ian Reid. Rgbd based dimensional decomposition residual network for 3d semantic scene completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. Cited on pages 16 and 85.

[LLJ+21] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021. Cited on pages 17 and 86.

[LLK+21] Sihaeng Lee, Janghyeon Lee, Byungju Kim, Eojindl Yi, and Junmo Kim. Patch-wise attention network for monocular depth estimation. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, volume 35, 2021. Cited on pages 14, 70, and 72.

[LLL18] Kyoungoh Lee, Inwoong Lee, and Sanghoon Lee. Propagating lstm: 3d pose estimation based on joint interdependency. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. Cited on pages 23 and 24.

[LLL+24] Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, et al. Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024. Cited on pages 16, 85, and 91.

[LLS+21] Jiefeng Li, Chengcheng Liu, Xiao Sun, Nan Yu, Yichen Wei, and Jianfeng Wang. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. Cited on pages 3 and 12.

[LLY+19] Jie Li, Yu Liu, Xia Yuan, Chunxia Zhao, Roland Siegwart, Ian Reid, and Cesar Cadena. Depth based semantic scene completion with position importance aware loss. *IEEE Robotics and Automation Letters (RA-L)*, 5(1):219–226, 2019. Cited on pages 16 and 85.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014. Cited on page 29.

[LMR+15] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6), 2015. Cited on pages 2 and 11.

[LS18] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. Cited on page 23.

[LSC+19] Zongmian Li, Jiri Sedlar, Justin Carpentier, Ivan Laptev, Nicolas Mansard, and Josef Sivic. Estimating 3d motion and forces of person-object interactions from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. Cited on pages 13 and 39.

[LSS+22] Zhi Li, Soshi Shimada, Bernt Schiele, Christian Theobalt, and Vladislav Golyanik. Mocapdeform: Monocular 3d human motion capture in deformable scenes. In *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, 2022. Cited on pages 8, 9, 10, 37, 39, and 51.

[LSSD23] Zhi Li, Shaoshuai Shi, Bernt Schiele, and Dengxin Dai. Test-time domain adaptation for monocular depth estimation. In *Proceedings of the IEEE International Conference on Robotics and Animation (ICRA)*. IEEE, 2023. Cited on pages 9, 10, 67, and 93.

[LVdC23] Ivan Lopes, Tuan-Hung Vu, and Raoul de Charette. Cross-task attention mechanism for dense multi-task learning. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2023. Cited on pages 17 and 86.

[LWL+22] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022. Cited on page 84.

[LWLJ22] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022. Cited on pages 14 and 70.

[LWWJ19] Zhi Li, Xuan Wang, Fei Wang, and Peilin Jiang. On boosting single-frame 3d human pose estimation via monocular videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. Cited on pages 8, 21, and 22.

[LWZ+19] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. Cited on page 22.

[LWZ+21] Tianyu Luan, Yali Wang, Junhao Zhang, Zhe Wang, Zhipeng Zhou, and Yu Qiao. Pc-hmr: Pose calibration for 3d human mesh recovery from 2d images/videos. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, volume 35, 2021. Cited on page 12.

[LXG22] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2022. Cited on page 91.

[LYC+23] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. Cited on pages 91, 92, and 93.

[LYW+23] Zhiqi Li, Zhiding Yu, Wenhai Wang, Anima Anandkumar, Tong Lu, and Jose M Alvarez. Fb-bev: Bev representation from forward-backward view transformations. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. Cited on page 84.

[MCL19] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. Cited on pages 11, 12, 13, 22, 23, 25, 27, 28, 29, 30, 31, 32, 33, 34, 35, 53, 105, and 109.

[Met21] Metashape. `http://www.agisoft.com`, 2021. Cited on pages 54 and 61.

[MGC+19] Aron Monszpart, Paul Guerrero, Duygu Ceylan, Ersin Yumer, and Niloy J Mitra. imapper: interaction-guided scene mapping from monocular videos. *ACM Transactions On Graphics (TOG)*, 38(4):1–15, 2019. Cited on pages 13 and 53.

[MGT+19] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. Cited on pages 39, 44, 56, and 61.

[MHK06] Thomas B Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding (CVIU)*, 104(2-3):90–126, 2006. Cited on pages 2 and 53.

[MHRL17] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. Cited on pages 2, 11, 12, 13, 23, 24, 28, 29, 37, 39, 51, and 53.

[MN17] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. Cited on pages 13, 28, 29, 37, 39, and 53.

[Mob20] Mobileye. Mobileye ces 2020 presentation. `https://youtu.be/HPWGFzqd7pI`, 2020. Accessed: 2025-03-06. Cited on page 83.

[MOT+21] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self-contact and human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. Cited on pages 13 and 39.

[MRC+17] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, 2017. Cited on pages 1, 2, 13, 22, 28, 29, 30, 33, 37, 39, and 53.

[MSM+18] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, 2018. Cited on pages 22, 23, and 28.

[MSM+20] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohammad Elgharib, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: Real-time multi-person 3d motion capture with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 39(4), 2020. Cited on pages 13, 22, 37, 39, 53, and 55.

[MSS+17] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017. Cited on pages 2, 13, 37, 39, 43, 51, 53, and 55.

[MST+21] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. Cited on pages 5, 6, 7, 16, 85, 86, 97, and 102.

[MWA18] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. Cited on pages 5 and 6.

[NIH+11] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 2011. Cited on page 6.

[NW06] Jorge Nocedal and Stephen J Wright. Nonlinear equations. *Numerical Optimization*, pages 270–302, 2006. Cited on page 60.

[NYD16] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016. Cited on pages 23, 37, and 39.

[ODM+25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. Cited on pages 85, 96, and 102.

[PBV21] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. Cited on page 12.

[PCG+19] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. Cited on pages 2, 11, 13, 41, 42, 44, 45, 46, 47, 52, 53, 54, 55, 56, 60, 61, and 62.

[PFGA19] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. Cited on pages 12, 37, and 39.

[PGC+17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Advances in Neural Information Processing Systems Workshops (NeurIPSW)*, 2017. Cited on page 60.

[PGJ+23] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. Cited on pages 16 and 86.

[PGLC15] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, 2015. Cited on pages 15 and 69.

[PHT23] Malte Prinzler, Otmar Hilliges, and Justus Thies. Diner: Depth-aware image-based neural radiance fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. Cited on pages 16 and 86.

[PJH23] Matt Pharr, Wenzel Jakob, and Greg Humphreys. *Physically based rendering: From theory to implementation*. MIT Press, 2023. Cited on page 94.

[PKS+19] Jhony K Pontes, Chen Kong, Sridha Sridharan, Simon Lucey, Anders Eriksson, and Clinton Fookes. Image2mesh: A learning framework for single image 3d reconstruction. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Springer, 2019. Cited on page 16.

[PLZ+24] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Hongwei Xie, Bing Wang, Li Liu, and Shanghang Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *Proceedings of the IEEE International Conference on Robotics and Animation (ICRA)*. IEEE, 2024. Cited on pages 6, 7, 16, 84, 85, 86, and 89.

[Pop07] Ronald Poppe. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding (CVIU)*, 108(1-2):4–18, 2007. Cited on pages 2 and 53.

[PTKY10] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010. Cited on pages 15 and 69.

[PZD18] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. Cited on page 11.

[PZDD17] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. Cited on pages 1, 2, 11, 12, 13, 23, 37, 39, and 53.

[PZZD18] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. Cited on pages 11, 13, 37, 39, 51, and 53.

[QSMG17] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. Cited on page 16.

[QYSG17] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. Cited on page 16.

[RBH+21] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. Cited on pages 2, 3, 12, 13, 14, 38, 39, 52, and 53.

[RC11] Radu Bogdan Rusu and Steve Cousins. 3d is here: Point cloud library (pcl). In *Proceedings of the IEEE International Conference on Robotics and Animation (ICRA)*. IEEE, 2011. Cited on page 5.

[RdCVB20] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, 2020. Cited on pages 16 and 85.

[RDCVB22] Luis Roldao, Raoul De Charette, and Anne Verroust-Blondet. 3d semantic scene completion: A survey. *International Journal of Computer Vision (IJCV)*, 130(8):1978–2005, 2022. Cited on page 85.

[RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115:211–252, 2015. Cited on page 31.

[REEG21] Christoph B Rist, David Emmerichs, Markus Enzweiler, and Dariu M Gavrila. Semantic scene completion using local deep implicit functions on lidar data. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 44(10):7205–7218, 2021. Cited on pages 85 and 91.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. Cited on page 88.

[RGH+20] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. Cited on pages 3, 12, 13, 14, 38, 39, 52, and 53.

[RKH+21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2021. Cited on page 85.

[RPKM21] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people with 3d representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. Cited on page 12.

[RSF18] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation learning for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. Cited on pages 12, 13, 37, 39, 51, and 53.

[RWS17] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. Cited on pages 12, 22, 23, 28, and 33.

[RWS19] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 42(5):1146–1161, 2019. Cited on pages 12, 28, and 33.

[SA07] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, pages 109–116, 2007. Cited on page 58.

[SAA+20] Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *ACM Transactions on Graphics (TOG)*, 40(1):1–15, 2020. Cited on pages 13, 37, 39, and 53.

[SBB10] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision (IJCV)*, 87:4–27, 2010. Cited on page 2.

[SBIK16] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A Kakadiaris. 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding (CVIU)*, 152:1–20, 2016. Cited on pages 11, 23, and 53.

[SCC+22] Chang Shu, Ziming Chen, Lei Chen, Kuan Ma, Minghui Wang, and Haibing Ren. Sidert: A real-time pure transformer architecture for single image depth estimation. *arXiv preprint arXiv:2204.13892*, 2022. Cited on pages 14 and 70.

[SCD+06] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1. IEEE, 2006. Cited on pages 4 and 6.

[SCN05] Ashutosh Saxena, Sung Chung, and Andrew Ng. Learning depth from single monocular images. *Advances in Neural Information Processing Systems (NeurIPS)*, 18, 2005. Cited on page 1.

[SF16] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. Cited on pages 1 and 6.

[SGL+22] Soshi Shimada, Vladislav Golyanik, Zhi Li, Patrick Pérez, Weipeng Xu, and Christian Theobalt. Hulc: 3d human motion capture with pose manifold sampling and dense contact guidance. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022. Cited on pages 8, 10, 37, 53, and 54.

[SGX+21] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3d human motion capture with physical awareness. *ACM Transactions on Graphics (TOG)*, 40(4), aug 2021. Cited on pages 12, 13, 38, 39, 43, and 52.

[SGXT20] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020. Cited on pages 12, 13, 38, 39, 43, 48, 51, 52, and 53.

[SHKF12] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2012. Cited on page 5.

[SHN+19] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. Cited on pages 40 and 41.

[SJMS17] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. Cited on page 55.

[ska] Skanect: 3d scanning. `https://skanect.occipital.com`. Cited on page 54.

[SLY15] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. Cited on page 40.

[SRS+12] Sanjay Saini, Dayang Rohaya Bt Awang Rambli, Suziah Bt Sulaiman, M Nordin B Zakaria, and Siti Rohkmah. Markerless multi-view human motion tracking using manifold model learning by charting. *Procedia Engineering*, 41:664–670, 2012. Cited on page 40.

[SRSZ13] Sanjay Saini, Dayang Rohaya Bt Awang Rambli, Suziah Bt Sulaiman, and M Nordin B Zakaria. Human pose tracking in low-dimensional subspace using manifold learning by charting. In *Proceedings of the International Conference on Signal and Image Processing Applications (ICSIPA)*, 2013. Cited on page 40.

[SS23] Taha Samavati and Mohsen Soryani. Deep learning-based 3d reconstruction: a survey. *Artificial Intelligence Review*, 56(9):9175–9219, 2023. Cited on page 5.

[SSLW17] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. Cited on pages 11 and 29.

[str] Structure sensor: 3d scanning, augmented reality and more. `https://structure.io/structure-sensor`. Cited on page 54.

[SVB+19] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. Cited on pages 28 and 40.

[SXW+18] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. Cited on pages 2, 11, 22, 27, 28, and 29.

[SYL+19] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, YiLi Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. Cited on pages 37 and 39.

[SYW+23] Xiaolong Shen, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Global-to-local modeling for video-based 3d human pose and shape estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. Cited on page 12.

[SYZ+17] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. Cited on pages 16 and 85.

[SZZ+21] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021. Cited on page 89.

[Tes21] Tesla. Tesla ai day 2021. `https://www.youtube.com/watch?v=j0z4FweCy4M`, 2021. Accessed: March 6, 2025. Cited on page 83.

[TGM+17] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and Jonathon Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017. Cited on page 2.

[THS+18] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. Cited on pages 15 and 69.

[TKS+16] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3d human pose with deep neural networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016. Cited on pages 13, 37, 39, and 53.

[TPMDS19] Alessio Tonioni, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Unsupervised domain adaptation for depth prediction from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 42(10):2396–2409, 2019. Cited on pages 15 and 69.

[TRA17] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. Cited on pages 13, 29, 37, 39, 51, and 53.

[TSW+23] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. Cited on pages 16, 83, 84, 85, and 91.

[TV17] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. Cited on pages 71, 75, and 107.

[USS+17] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, 2017. Cited on page 76.

[Vic] Vicon blade. `https://www.vicon.com/`. Cited on page 45.

[VL19] Márton Véges and András Lőrincz. Absolute human pose estimation with depth prediction network. In *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019. Cited on pages 11, 12, 22, 23, 28, 29, 30, 32, 34, and 109.

[VVGG+20] Simon Vandenhende, Wouter Van Gansbeke, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. MTI-Net: Multi-Scale Task Interaction Networks for Multi-Task Learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. Cited on pages 17, 85, and 86.

[WBP18] Markus Wulfmeier, Alex Bewley, and Ingmar Posner. Incremental adversarial domain adaptation for continually changing environments. In *Proceedings of the IEEE International Conference on Robotics and Animation (ICRA)*. IEEE, 2018. Cited on pages 15 and 70.

[WBSS04] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 13(4):600–612, 2004. Cited on page 90.

[WC10] Xiaolin Wei and Jinxiang Chai. Videomocap: Modeling physically realistic human motion from monocular video sequences. *ACM Transactions on Graphics (TOG)*, 29(4), 2010. Cited on pages 37 and 39.

[WCK+25] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. Cited on page 102.

[WCR+22] Zhe Wang, Liyan Chen, Shaurya Rathore, Daeyun Shin, and Charless Fowlkes. Geometric pose affordance: 3d human pose with scene constraints. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, 2022. Cited on pages 13, 14, 45, 46, 47, 48, 54, and 109.

[WDH+21] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. Cited on pages 5, 15, and 69.

[WFVGD22] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. Cited on pages 15, 68, 69, 70, 77, 78, and 79.

[WLR22] Bastian Wandt, James J. Little, and Helge Rhodin. Elepose: Unsupervised 3d human pose estimation by predicting camera elevation and learning normalizing flows on 2d poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. Cited on pages 11 and 51.

[WLZ+24] Yuqun Wu, Jae Yong Lee, Chuhang Zou, Shenlong Wang, and Derek Hoiem. Monopatchnerf: Improving neural radiance fields with patch-based monocular guidance. *arXiv preprint arXiv:2404.08252*, 2024. Cited on pages 16 and 86.

[WMAP+21] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. Cited on pages 6, 15, and 70.

[WMM+21] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021. Cited on page 12.

[WR19] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. Cited on page 11.

[WRKS16] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. Cited on page 55.

[WSF20] Zhe Wang, Daeyun Shin, and Charless Fowlkes. Predicting camera viewpoint improves cross-dataset generalization for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, 2020. Cited on pages 45, 46, 47, 48, and 109.

[WSK+15] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. Cited on page 6.

[WSL+20] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. Cited on pages 15 and 69.

[WTNT20] Shun-Cheng Wu, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scfusion: Real-time incremental scene reconstruction with semantic completion. In *Proceedings of the International Conference on 3D Vision (3DV)*. IEEE, 2020. Cited on pages 16 and 85.

[WTZ+21] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. Deep 3d human pose estimation: A review. *Computer Vision and Image Understanding (CVIU)*, 210:103225, 2021. Cited on pages 3 and 105.

[WXX+21] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. Cited on page 39.

[WY21] Zhenzhen Weng and Serena Yeung. Holistic 3d human and scene mesh estimation from single view images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. Cited on pages 13, 14, 52, and 53.

[WYRC23] Felix Wimbauer, Nan Yang, Christian Rupprecht, and Daniel Cremers. Behind the scenes: Density fields for single view reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. Cited on pages 6 and 85.

[WZX+23] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. Cited on pages 16, 83, 84, 85, and 91.

[WZZ+23] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. Cited on pages 16 and 85.

[XGD+17] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. Cited on pages 31 and 33.

[XGF16] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016. Cited on page 4.

[XOWS18] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. Cited on pages 17, 85, and 86.

[XXG+20] Lan Xu, Weipeng Xu, Vladislav Golyanik, Marc Habermann, Lu Fang, and Christian Theobalt. Eventcap: Monocular 3d capture of high-speed human motions using an event camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Cited on page 53.

[XZH+20] Feng Xue, Guirong Zhuo, Ziyuan Huang, Wufei Fu, Zhuoyue Wu, and Marcelo H Ang. Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020. Cited on pages 15, 70, 77, 78, and 79.

[XZV+22] Xiaogang Xu, Hengshuang Zhao, Vibhav Vineet, Ser-Nam Lim, and Antonio Torralba. Mtformer: Multi-task learning via transformer and cross-task reasoning. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022. Cited on pages 17, 85, and 86.

[YGD+22] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. New crfs: Neural window fully-connected crfs for monocular depth estimation. *arXiv preprint arXiv:2203.01502*, 2022. Cited on pages 70, 71, 72, 77, 78, 80, and 81.

[YGL+21] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, volume 35, 2021. Cited on pages 16 and 85.

[YHL+23] Hao Yang, Lanqing Hong, Aoxue Li, Tianyang Hu, Zhenguo Li, Gim Hee Lee, and Liwei Wang. Contranerf: Generalizable neural radiance fields for synthetic-to-real novel view synthesis via contrastive learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. Cited on pages 16 and 86.

[YIK+16] Hashim Yasin, Umar Iqbal, Bjorn Kruger, Andreas Weber, and Juergen Gall. A dual-source approach for 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. Cited on page 28.

[YKH+24a] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. Cited on pages 85, 101, and 102.

[YKH+24b] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. Cited on pages 85, 88, 90, 92, 96, 101, and 102.

[YLL+18] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. Cited on page 1.

[YLS+23] Jiawei Yao, Chuming Li, Keqiang Sun, Yingjie Cai, Hao Li, Wanli Ouyang, and Hongsheng Li. Ndc-scene: Boost monocular 3d semantic scene completion in normalized device coordinates space. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2023. Cited on page 85.

[YLSY19] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. Cited on pages 14 and 70.

[YLZ21] Fuming You, Jingjing Li, and Zhou Zhao. Test-time batch statistics calibration for covariate shift. *arXiv preprint arXiv:2110.04065*, 2021. Cited on pages 15 and 69.

[YOW+18] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. Cited on pages 23, 24, 28, 37, and 39.

[YPN+22] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 2022. Cited on pages 16 and 86.

[YS20] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Cited on pages 15 and 69.

[YWS+21] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. Cited on page 39.

[YWvdW+21] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. Cited on pages 15 and 69.

[YYTK21] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. Cited on pages 85 and 86.

[YYYH21] Hao-Wei Yeh, Baoyao Yang, Pong C Yuen, and Tatsuya Harada. Sofa: Source-data-free feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021. Cited on pages 15 and 69.

[YZH+22] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. Cited on pages 38 and 52.

[ZBSL17] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. Cited on pages 1, 12, 14, and 70.

[ZCX+19] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. Cited on pages 14 and 70.

[ZDM23] Junbo Zhang, Runpei Dong, and Kaisheng Ma. Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. Cited on page 84.

[ZF18] Yinda Zhang and Thomas Funkhouser. Deep depth completion of a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. Cited on page 1.

[ZHH+21] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular real-time full body capture with inter-part correlations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. Cited on pages 13 and 53.

[ZHN+20] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Cited on page 62.

[ZHS+17a] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. Cited on pages 2, 37, and 39.

[ZHS+17b] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Weakly-supervised transfer for 3d human pose estimation in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, volume 3, 2017. Cited on pages 11, 23, and 24.

[ZHW20] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Cited on page 39.

[ZL21] Aurick Zhou and Sergey Levine. Training on test data with bayesian adaptation for covariate shift. *arXiv preprint arXiv:2109.12746*, 2021. Cited on pages 15 and 69.

[ZLLD21] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. Cited on pages 16 and 86.

[ZMS18] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. Cited on pages 12, 13, 22, 38, 39, and 53.

[ZOL+20] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3d human mesh regression with dense correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Cited on page 12.

[ZPJ+20] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020. Cited on pages 13, 52, and 53.

[ZPK18] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. Cited on pages 5 and 60.

[ZPT+19] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. Cited on page 28.

[ZPZS21] Andrei Zanfir, Alin-Ionut Popa, Mihai Zanfir, and Cristian Sminchisescu. Thundr: Transformer-based 3d human reconstruction with markers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. Cited on pages 4 and 12.

[ZSL+19] Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, and Youliang Yan. Exploiting temporal consistency for real-time video depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. Cited on pages 14 and 70.

[ZSZ+16] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. Deep kinematic pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. Cited on page 53.

[ZWC+23] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 56(1):1–37, 2023. Cited on pages 2, 11, and 23.

[ZWW+25] Xiaoyu Zhou, Jingqi Wang, Yongtao Wang, Yufei Wei, Nan Dong, and Ming-Hsuan Yang. Occgs: Zero-shot 3d occupancy reconstruction with semantic and geometric-aware gaussian splatting. *arXiv preprint arXiv:2502.04981*, 2025. Cited on pages 16 and 86.

[ZYC+20] Yuliang Zou, Jimei Yang, Duygu Ceylan, Jianming Zhang, Federico Perazzi, and Jia-Bin Huang. Reducing footskate in human motion reconstruction with ground contact constraints. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020. Cited on pages 13 and 39.

[ZYKW18] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. Cited on pages 15 and 69.

[ZYW+23] Chubin Zhang, Juncheng Yan, Yi Wei, Jiaxin Li, Li Liu, Yansong Tang, Yueqi Duan, and Jiwen Lu. Occnerf: Self-supervised multi-camera occupancy prediction with neural radiance fields. *arXiv preprint arXiv:2312.09243*, 2023. Cited on pages 16, 85, 86, and 89.

[ZZB+21] Siwei Zhang, Yan Zhang, Federica Bogo, Pollefeys Marc, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. Cited on pages 13, 14, 38, 39, 40, 45, 46, 47, 50, 52, 53, 54, and 62.

[ZZD23] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. Cited on pages 7, 91, 92, and 93.

[ZZM+20] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. Place: Proximity learning of articulation and contact in 3d environments. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2020. Cited on page 62.

[ZZP+18] Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 41(4):901–914, 2018. Cited on pages 28 and 29.