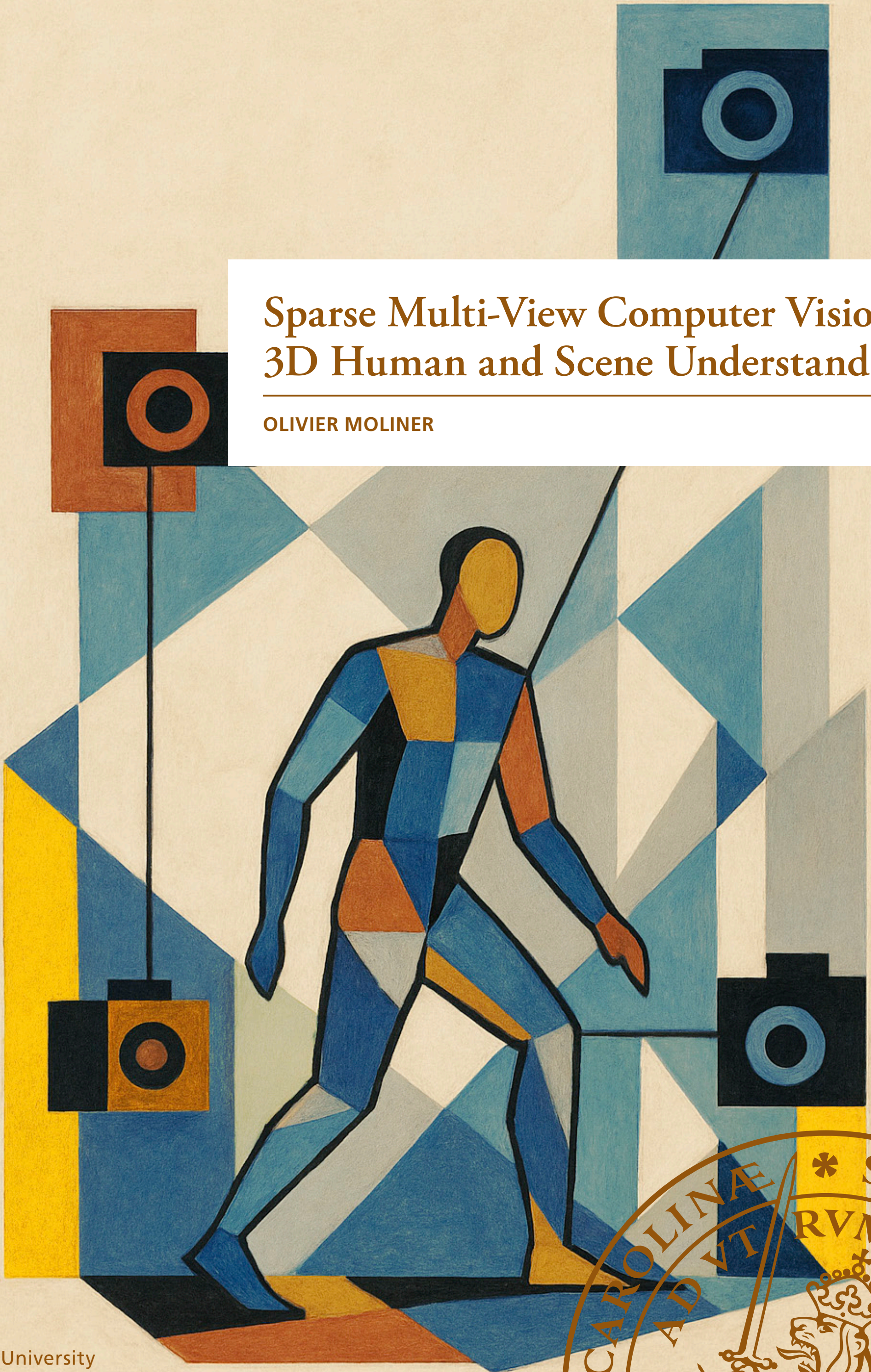
Sparse Multi-View Computer Vision for 3D Human and Scene Understanding

Moliner, Olivier

2025

– CENTRUM SCIENTIARUM MATHEMATICARUM –



Sparse Multi-View Computer Vision for 3D Human and Scene Understanding

OLIVIER MOLINER

Lund University
Faculty of Engineering
Centre for Mathematical Sciences
Mathematics

Sparse Multi-View Computer Vision for 3D Human and Scene Understanding

by Olivier Moliner

Thesis for the degree of Doctor of Philosophy in Engineering

Thesis advisors:
Prof. Kalle Åström, Dr. Sangxia Huang, Assoc. Prof. Viktor Larsson,
Prof. Fredrik Tufvesson, Andrej Petef

Faculty opponent:
Prof. Helge Rhodin, Bielefeld University, Germany

To be presented, with the permission of the Faculty of Engineering at Lund University, for public criticism in the Hörmander lecture hall at the Centre for Mathematical Sciences on Friday, the 10th of October 2025 at 13:15.

DOKUMENTDATABLAD enl SIS 61 41 21

Organization
LUND UNIVERSITY
Centre for Mathematical Sciences
Box 118
SE–221 00 LUND
Sweden

Document name
Doctoral thesis

Date of presentation
2025-10-10

Sponsoring organization
Sony, WASP

Author(s)
Olivier Moliner

Title and subtitle
Sparse Multi-View Computer Vision for 3D Human and Scene Understanding

Abstract

Perceiving and understanding human motion is a fundamental problem in computer vision, with diverse applications encompassing sports analytics, healthcare monitoring, entertainment, and intelligent interactive systems. Multi-camera systems, by capturing multiple viewpoints simultaneously, enable robust tracking and reconstruction of human poses in 3D, overcoming limitations of single-view approaches. This thesis addresses key bottlenecks encountered when designing and deploying multi-camera systems for 3D human and scene understanding beyond controlled laboratory settings.

Paper I introduces a human-pose-based approach to extrinsic camera calibration that leverages naturally occurring human motion in the scene. By incorporating a 3D pose likelihood model in kinematic chain space and a distance-aware confidence-weighted reprojection loss, we enable accurate wide-baseline calibration without calibration equipment. This allows for rapid deployment and reconfiguration of multi-camera systems without requiring technical expertise.

The reliance on large labeled datasets presents a significant obstacle to the widespread adoption of action recognition systems. In Paper II we propose a self-supervised learning framework for skeleton-based action recognition. We adapted Bootstrap Your Own Latent (BYOL) for 3D human pose sequence representation. Our contributions include multi-viewpoint sampling that leverages existing multi-camera data, and asymmetric augmentation pipelines bridging the domain shift gap when fine-tuning the network for downstream tasks. This self-supervised method reduces the need for labeled data, shortening development time for new applications.

Paper III focuses on robust 3D human pose reconstruction, particularly in challenging real-world scenarios. Triangulation-based methods struggle in occluded or sparsely-covered scenes. We designed an encoder-decoder Transformer model that regresses 3D human poses from multi-view 2D pose sequences, and introduced a biased attention mechanism that leverages geometric relationships between views and detection confidence scores. Our approach enables robust reconstruction of 3D human poses under heavy occlusion and when few input views are available.

In Paper IV, we tackle open-vocabulary 3D object detection from sparse multi-view RGB data. Our approach builds on pre-trained, off-the-shelf 2D networks and does not require retraining. We lift 2D detections into 3D via monocular depth estimation, followed by multi-view feature consistency optimization and 3D fusion of sparse proposals. Our experiments show that this approach can produce comparable results to state-of-the-art methods in the densely sampled setting while significantly outperforming the state-of-the-art for instances with sparse-views.

Key words
Multi-view Geometry; Extrinsic Camera Calibration; Multi-camera System; 3D Human Pose Estimation; Skeleton-based Action Recognition; Self-supervised Learning; 3D Object Detection; 3D Scene Understanding

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title
1404-0034. Doctoral Theses in Mathematical Sciences

ISBN
978-91-8104-604-5 (print)
978-91-8104-605-2 (pdf)

Recipient's notes

Number of pages
xviii+172

Price

Security classification

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources the permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature ____________________ Date 2025-08-28

Sparse Multi-View Computer Vision for 3D Human and Scene Understanding

by Olivier Moliner

LUND
UNIVERSITY

Cover illustration: A Cubist-inspired digital painting symbolizing human motion and sparse camera perspectives in 3D scene understanding, created by the author in collaboration with ChatGPT.

Centre for Mathematical Sciences
Lund University
Box 118
SE-221 00 Lund
Sweden
www.maths.lu.se

Doctoral Theses in Mathematical Sciences 2025:3
ISSN: 1404-0034
ISBN: 978-91-8104-604-5 (print)
ISBN: 978-91-8104-605-2 (electronic)
LUTFMA-1089-2025

Printed in Sweden by Media-Tryck, Lund University, Lund 2025

NORDIC SWAN ECOLABEL
Printed matter
3041 0903

Media-Tryck is a Nordic Swan Ecolabel
certified provider of printed material.
Read more about our environmental
work at www.mediatryck.lu.se

MADE IN SWEDEN

Avant donc que d'écrire, apprenez à penser. [...]
Ce que l'on conçoit bien s'énonce clairement,
Et les mots pour le dire arrivent aisément. [...]
Hâtez-vous lentement, et, sans perdre courage,
Vingt fois sur le métier remettez votre ouvrage:
Polissez-le sans cesse et le repolissez;
Ajoutez quelquefois, et souvent effacez.

Nicolas Boileau, *L'Art poétique*, 1669

Abstract

Perceiving and understanding human motion is a fundamental problem in computer vision, with diverse applications encompassing sports analytics, healthcare monitoring, entertainment, and intelligent interactive systems. Multi-camera systems, by capturing multiple viewpoints simultaneously, enable robust tracking and reconstruction of human poses in 3D, overcoming limitations of single-view approaches. This thesis addresses key bottlenecks encountered when designing and deploying multi-camera systems for 3D human and scene understanding beyond controlled laboratory settings.

Paper I introduces a human-pose-based approach to extrinsic camera calibration that leverages naturally occurring human motion in the scene. By incorporating a 3D pose likelihood model in kinematic chain space and a distance-aware confidence-weighted reprojection loss, we enable accurate wide-baseline calibration without calibration equipment. This allows for rapid deployment and reconfiguration of multi-camera systems without requiring technical expertise.

The reliance on large labeled datasets presents a significant obstacle to the widespread adoption of action recognition systems. In Paper II we propose a self-supervised learning framework for skeleton-based action recognition. We adapted Bootstrap Your Own Latent (BYOL) for 3D human pose sequence representation. Our contributions include multi-viewpoint sampling that leverages existing multi-camera data, and asymmetric augmentation pipelines bridging the domain shift gap when fine-tuning the network for downstream tasks. This self-supervised method reduces the need for labeled data, shortening development time for new applications.

Paper III focuses on robust 3D human pose reconstruction, particularly in challenging real-world scenarios. Triangulation-based methods struggle in occluded or sparsely-covered scenes. We designed an encoder-decoder Transformer model that regresses 3D human poses from multi-view 2D pose sequences, and introduced a biased attention mechanism that leverages geometric relationships between views and detection confidence scores. Our approach enables robust reconstruction of 3D human poses under heavy occlusion and when few input views are available.

In Paper IV, we tackle open-vocabulary 3D object detection from sparse multi-view RGB data. Our approach builds on pre-trained, off-the-shelf 2D networks and does not require retraining. We lift 2D detections into 3D via monocular depth estimation, followed by multi-view feature consistency optimization and 3D fusion of sparse proposals. Our experiments show that this approach can

produce comparable results to state-of-the-art methods in the densely sampled setting while significantly outperforming the state-of-the-art for instances with sparse-views.

Popular Summary

Human motion has fascinated scholars and artists since ancient times. The advent of photography and film, allowing to capture movement accurately, paved the way for modern scientific study of motion. Modern computer systems can now identify where a person's body parts are located in images, creating simple stick-figure representations that track how people move, enabling automatic analysis. But a single camera viewpoint provides limited information: it is impossible to tell how far away objects are, and objects or other people may block the view. With several cameras capturing a scene from different angles, we can estimate true 3D information, much like how our two eyes work together to give us depth perception. This thesis proposes methods for building practical, robust systems that can perceive and understand human activities and scenes in 3D using multiple cameras.

Before multiple cameras can work together, their exact positions and orientations must be known. This can be done by *calibrating* the cameras, a process often involving special equipment, like checkerboard patterns, waved through the scene following specific procedures, which often requires a trained operator to achieve good accuracy. This can be long and expensive, making it impractical for scenarios requiring fast installation. In this thesis, we present a new calibration method that uses the people moving in the scene as calibration patterns, enabling faster installation without equipment: all it takes is for someone to walk around the scene for a while something that might happen naturally in a new installation.

In controlled laboratory conditions, perfectly calibrated cameras would allow flawless reconstruction of human poses in 3D. But real-world environments present significant challenges. Furniture, other people, or even the environment itself may hide parts of a person, lighting conditions may vary, and the initial position estimates can be noisy or inaccurate. Simply combining these imperfect observations can result in distorted or absurd body poses. To reconstruct accurate 3D human poses even under challenging conditions, we trained a neural network to pay attention to the geometric relationships between views and to the movements of 2D poses over time. It learns to create coherent and accurate 3D human posture and motion even when some body parts are not visible to any of the cameras, filling in the gaps with realistic predictions based on how people naturally move.

Once we have accurate 3D movement sequences, we can teach computers to recognize different activities, like ”walking” or ”doing push-ups”. Training computers for *activity recognition* usually requires large amounts of *labeled data*, where humans manually mark and name each activity in thousands of motion sequences. This is labour-intensive and very expensive. What's more, recognizing a new activity requires collecting and labeling new data from scratch. On the other hand, just collecting *unlabeled* movement data is easy and inexpensive. We developed a *self-supervised learning* method that allows computers to learn meaningful patterns from large amounts of unlabeled movement data. For example, it learns that an action remains fundamentally the same even when viewed from different angles, or when performed at different speeds. After this initial learning phase, called *pre-training*, the system is very adaptable: it can quickly learn to recognize new activities using just a few labeled examples, significantly reducing the time and cost of training.

Truly understanding human behaviour requires more than just tracking body movement. We interact constantly with our environment, and our movements take on specific meaning depending on the context: standing in front of a refrigerator suggests that you will open it; moving towards a chair indicates that you are about to sit down. Without seeing the surrounding objects, these movements could mean almost anything. We developed a method for localizing objects in 3D using just few camera views and everyday language descriptions, like ”a blue bowl” or ”an office chair”. This enables the system to find objects it has never been trained to recognize, providing rich environmental context for human activity.

Together, these contributions help move multi-camera 3D perception systems from the lab to the real world, making them more accessible, efficient, and effective for diverse applications such as sports analysis, safety monitoring, entertainment, and beyond.

Résumé en Français

Le mouvement humain a fasciné scientifiques, philosophes et artistes depuis l'Antiquité. L'avènement de la photographie et du cinéma, permettant de capturer les postures avec précision, a ouvert la voie à l'étude scientifique moderne du mouvement. Les systèmes d'intelligence artificielle (IA) peuvent désormais identifier la position des parties du corps humain à partir d'images, créant des représentations simplifiées en forme de "bonhomme allumette" correspondant à des points clés anatomiques, permettant une analyse automatique. Mais un seul point de vue ne fournit que des informations limitées : il est impossible de déterminer à quelle distance se trouvent les objets, et la vue peut être partiellement bloquée. En utilisant plusieurs vues de la même scène sous différents angles, il est possible d'obtenir des informations en 3D, de la même façon que nos deux yeux travaillent ensemble pour nous donner la perception de la profondeur. Cette thèse propose des méthodes pour concevoir des systèmes pratiques et robustes capables de percevoir et de comprendre les activités humaines et les scènes en 3D en utilisant plusieurs caméras.

Pour que plusieurs caméras puissent fonctionner ensemble, leurs positions et orientations respectives doivent être connues exactement. Cela peut se faire en *calibrant* les caméras, un processus qui requiert un équipement spécialisé, comme des motifs en damier, déplacés à travers la scène selon des procédures spécifiques, ce qui nécessite souvent un opérateur expert pour obtenir une bonne précision. Cela peut être long et coûteux, rendant impratiques les applications nécessitant une installation rapide. Dans cette thèse, nous présentons une nouvelle méthode de calibrage qui utilise les personnes se déplaçant dans la scène comme motifs de calibrage, permettant une installation plus rapide sans équipement: il suffit que quelqu'un se promène dans la scène pendant un moment.

Dans des conditions de laboratoire contrôlées, des caméras parfaitement calibrées permettraient une reconstruction parfaite des poses humaines en 3D. Mais les environnements réels présentent des défis importants. Les meubles, d'autres personnes, ou même l'environnement lui-même peuvent occulter des parties d'une personne, les conditions d'éclairage peuvent varier, et les estimations de posture initiales peuvent être bruitées ou inexactes. Simplement combiner ces observations imparfaites aboutit souvent à des postures 3D déformées ou absurdes. Pour reconstruire des poses humaines 3D précises même dans des conditions difficiles, nous avons entraîné un réseau de neurones à prêter attention aux relations géométriques entre les vues et à l'évolution des poses 2D au fil du temps. Il apprend à recréer des postures et mouvements humains 3D cohérents et précis même lorsque certaines parties du corps ne sont visibles par

aucune des caméras, comblant les lacunes avec des prédictions réalistes basées sur la façon dont les gens se déplacent naturellement.

Une fois que nous avons des séquences de mouvement 3D précises, nous pouvons enseigner aux ordinateurs à reconnaître différentes activités, comme ”marcher” ou ”faire des pompes”. Entraîner des ordinateurs pour la reconnaissance d'activités nécessite habituellement de grandes quantités de données annotées, où des humains marquent et étiquettent manuellement chaque activité dans des milliers de séquences de mouvement. C'est intensif en main-d'œuvre et très coûteux. De plus, reconnaître une nouvelle activité nécessite de collecter et d'étiqueter de nouvelles données. D'autre part, simplement collecter des données de mouvement non étiquetées est facile et peu coûteux. Nous avons développé une méthode d'apprentissage auto-supervisé qui permet aux ordinateurs d'apprendre des représentations significatives à partir de grandes quantités de données de mouvement non étiquetées. Par exemple, elle permet d'apprendre qu'une action reste fondamentalement la même alors qu'elle est vue sous différents angles, ou lorsqu'elle est exécutée à différentes vitesses. Après cette phase d'apprentissage initial, appelée pré-entraînement, le système est très adaptable: il peut être entraîné à reconnaître de nouvelles activités en utilisant seulement quelques exemples étiquetés, réduisant considérablement le temps et le coût d'entraînement.

Comprendre le comportement humain nécessite plus que de simplement suivre le mouvement du corps. Nous interagissons constamment avec notre environnement, et nos mouvements prennent un sens spécifique selon le contexte: vous tenir debout devant un réfrigérateur suggère que vous allez l'ouvrir; vous diriger vers une chaise indique que vous êtes sur le point de vous asseoir. Sans voir les objets environnants, ces mouvements pourraient signifier presque n'importe quoi. Nous avons développé une méthode pour localiser des objets en 3D en utilisant seulement quelques points de vue ainsi que des descriptions en langage courant, comme ”un bol bleu” ou ”une chaise de bureau”. Cela permet au système de trouver des objets qu'il n'a jamais été entraîné à reconnaître, fournissant un contexte environnemental riche pour l'activité humaine.

Ensemble, ces contributions aident à faire passer les systèmes de perception 3D multi-caméras du laboratoire au monde réel, les rendant plus accessibles, efficients et adaptables à diverses applications telles que l'analyse sportive, la surveillance de sécurité, le divertissement, et au-delà.

List of Publications

This thesis is based on the following publications, referred to by their Roman numerals. They are reproduced and included in this thesis with the permission of their respective publishers. The author's contributions to each paper is listed below.

I **Better Prior Knowledge Improves Human-Pose-Based Extrinsic Camera Calibration**
Olivier Moliner, Sangxia Huang, Kalle Åström
25th International Conference on Pattern Recognition (ICPR), 2021.

Author's contributions: The idea was jointly discussed between all co-authors. OM implemented the method and performed the experiments. OM wrote most of the paper with feedback and contributions from KÅ and SH.

II **Bootstrapped Representation Learning for Skeleton-Based Action Recognition**
Olivier Moliner, Sangxia Huang, Kalle Åström
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2022.
Author's contributions: OM and SH suggested the project. OM developed the method with help from KÅ and SH, implemented the code and performed the experiments. OM wrote most of the paper with feedback from KÅ and SH.

III **Geometry-Biased Transformer for Robust Multi-View 3D Human Pose Reconstruction**
Olivier Moliner, Sangxia Huang, Kalle Åström
IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG), 2024.
Author's contributions: OM and SH proposed the project. OM developed the method with help from KÅ and SH, implemented the code and performed the experiments. OM wrote most of the paper with feedback from KÅ and SH.

IV **Sparse Multiview Open-Vocabulary 3D Detection**
Olivier Moliner, Viktor Larsson, Kalle Åström
Accepted for publication in *IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025.
Author's contributions: OM proposed the idea and developed the method with input from KÅ and VL. OM implemented the code and performed the experiments. The paper was written jointly by all co-authors.

Acknowledgements

This thesis marks the end of a seven-year journey that began when I never imagined a PhD was still within reach. I owe its completion to many people who have supported, challenged, and guided me along the way. First and foremost, I would like to thank my main supervisor, Kalle Åström, for his unfailing positivity and his generosity with time and ideas. His encouragement and calm guidance have been invaluable throughout these years. I am also grateful to my industrial supervisor Sangxia Huang, whose sharp insights and high expectations helped me grow during this journey. Thanks to Viktor Larsson, whose support, advice and late-night LaTeX-fu helped me push through the last stages of my project. My thanks also go to my co-supervisor Fredrik Tufvesson for the many stimulating discussions in the early stages of my doctoral studies as I was finding my research direction.

I am very grateful to Peter Karlsson, who gave me the chance to embark on this PhD when I thought that door had closed. His support in the early stages set me on this path. I would also like to thank Andrej Petef, who played a crucial role in making this industrial PhD possible, and who consistently supported me and encouraged me to pursue my academic goals.

To my colleagues at Sony, past and present, thank you for the support, the stimulating conversations, and the many moments that made work enjoyable. I am also grateful to the Computer Vision and Machine Learning group and my fellow PhD students. Even though I did not spend as much time at the department as I wished, I always felt welcomed. Special thanks to Martin and Malte for their pleasant company while sharing offices.

Many thanks to Kalle, Viktor, Andrej, Roberto and David for their thorough proofreading of this thesis and their constructive feedback that significantly improved the final manuscript.

On a more personal note, I want to thank my parents, who gave me unwavering support throughout my studies. They encouraged my curiosity from an early age and always believed in me, and for that I will always be grateful.

Most of all, I owe my deepest gratitude to my wife Henrietta. Without you, this thesis would never have been finished. Your support, patience, and love carried me through the hardest stretches, and your companionship made the good moments even better. As you now approach the completion of your own thesis, I hope I can offer you the same strength and encouragement you have given me.

Finally, to our children, Mathilde and Félix: you arrived during this PhD and changed everything. You reminded me of what really matters, and you brought joy and perspective in ways no research result ever could. I really hope we didn't scare you off from academia!

Funding

This work was financially supported by Sony and by the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg (KAW) Foundation.

Contents

Chapter 1

Introduction

Perceiving and understanding human motion is a longstanding challenge in research, with roots extending back to the earliest scientific studies of anatomy. Accurate 3D motion capture and analysis has become an essential tool for many applications such as sports analytics [5, 134, 145], healthcare [14, 76, 104, 124], and 3D content production [125, 157]. This thesis presents mehods for building and deploying versatile, accessible and scalable systems for analyzing human motion and understanding scenes in 3D from sparse multi-camera systems.

Marker-based motion capture systems such as Vicon [4], Qualisys [3] or OptiTrack [2] are currently the gold standard for precise motion analysis [139]. Widely used in biomechanics research, film production, and clinical gait analysis, these systems use reflective markers placed on the body and tracked by infrared cameras, enabling them to track movement with millimeter precision. However, they require expensive specialized equipment and controlled environments. Wearable inertial sensors provide an alternative that is more flexible and less expensive, as they do not require a capture rig. However, placing markers or sensors on the subjects is intrusive, time-consuming and exact device placement is difficult to reproduce between capture sessions [49, 101]. Measuring human movement should ideally be non-invasive and allow to capture subjects in their natural environment, such as workplaces, sport fields or public places, without encumbering their movements and without preparation.

Recent advances in deep learning have enabled markerless human pose estimation from images [15, 80, 133, 137, 140, 156], paving the way for motion capture using only cameras, which are ubiquitous, inexpensive, and unobtrusive. However, monocular approaches to human pose estimation face fundamental limitations that restrict their applicability in many practical scenarios. The most significant limitation is the inherent depth ambiguity arising from perspective projection: infinitely many 3D body poses can produce identical 2D projections [168]. Occlusion presents another critical challenge for single-view systems. Self-occlusions occur naturally during human movement as body parts move in front of each-other, while environmental occlusions arise from furniture, equipment, or other people in the scene. When body joints are not visible in the image, monocular methods must rely on learned priors about human anatomy and motion, which may fail for unusual poses or activities not well-represented in training data. Furthermore, single-camera systems provide only limited spatial coverage of the capture area. Subjects moving outside the camera's field of view are completely lost, and even within the visible area, pose estimation accuracy typically degrades with distance from the camera center. This spatial limitation severely constrains the types of activities that can be monitored and analyzed, particularly in large environments or scenarios requiring freedom of movement.

Multi-camera systems offer compelling solutions to many single-view limitations by capturing the scene from multiple viewpoints simultaneously. The fundamental geometric advantage lies in depth resolution: with calibrated cameras observing the same 3D scene, accurate depth information can be recovered by triangulation without the ambiguities inherent in monocular estimation. This allows reconstruction of poses in absolute world coordinates, enabling applications that require precise spatial measurements. The spatial redundancy provided by multiple cameras also significantly improves robustness to occlusions. Body parts occluded in one view may remain visible in others, allowing the system to reconstruct 3D body poses even under challenging conditions. This redundancy also enables more reliable pose estimation, as inconsistent observations across views can be identified and corrected for.

While multi-view human pose estimation has achieved impressive results, the focus has mainly been on developing accurate methods in controlled environments. Most academic research in multi-view pose estimation has been conducted using carefully designed setups to maximize system performance, such as optimal lighting conditions, uncluttered backgrounds, and precisely calibrated camera systems positioned to provide maximal spatial coverage with minimal occlusions. The Human3.6M [63] and CMU Panoptic [67] datasets, which serve as standard benchmarks in the field, exemplify this controlled approach with their well-lit studio setups and marker-based motion capture ground truth. Some high-end systems, such as Hawk-Eye [1], have taken multi-camera systems out of the lab for accurate 3D motion analysis, but they are very expensive, their installation is time-consuming and requires expertise, and they are designed to perform extremely well in very well-defined settings.

In this thesis, we present methods for developing and deploying cost-effective and versatile multi-camera systems for human and scene understanding. They enable fast installation and calibration without requiring dedicated calibration equipment or expertise, flexible configurations, and robust performance across diverse environments.

1.1 Challenges and Research Questions

The work presented in this thesis was conducted during the design and implementation of MCS, a multi-camera 3D perception platform developed at Sony's R&D center in Lund, which we present in Chapter 2. When designing the MCS system, our core objective was to create a system that was not only capable but also practical, robust enough for deployment in challenging environments, and adaptable to diverse applications. The implementation and deployment of MCS revealed challenges and inspired ideas that existing approaches had not adequately addressed. These directly motivated the research questions that structure this thesis.

A prerequisite for extracting 3D information from multiple camera views is to calibrate the cameras to establish their relative poses. Traditional multi-view systems require expert calibration using specialized equipment such as checkerboard patterns or structured light projectors. This process is time-consuming, requires trained personnel, and must be repeated whenever cameras are moved. Exploring new ideas often involves evaluating different camera situations in various environments and collaborating remotely with domain experts who are not always tech-savvy. This leads to our first question:

RQ1: *How can extrinsic camera calibration be made more accessible while maintaining accuracy?*

This challenge motivated Paper I, where we developed a human-pose-based calibration method that uses naturally occurring human motion as calibration patterns, enabling rapid deployment without specialized equipment or expertise.

Real-world deployment environments present numerous challenges that are difficult to anticipate in laboratory settings. Room layouts dictate possible camera placement, often resulting in suboptimal viewing angles and coverage gaps. The availability of power outlets and the practicality of running cables constrain system architecture, frequently forcing compromises between spatial coverage and installation feasibility. Lighting conditions vary throughout the day and may cause challenging conditions such as backlighting, harsh shadows, or insufficient illumination. These issues may produce noisy or missing 2D body part detections. Traditional triangulation-based reconstruction methods fail catastrophically in these conditions. The question emerged:

RQ2: *How can we achieve robust 3D human pose reconstruction with sparse views and in challenging environments?*

Paper III presents a learning-based approach that treats multi-view 3D reconstruction as a regression problem, leveraging geometric constraints and temporal continuity to handle missing and noisy observations.

Accurate 3D human pose estimation enables further processing and analysis. It is for example possible to classify the activities of the persons moving in the scene from sequences of 3D poses. Traditional supervised learning approaches demand extensive datasets with labeled motion sequences. This means that, even for exploring potential ideas, a considerable amount of time and resources need to be spent in gathering, curating and annotating data, before the idea can be evaluated. This is both costly and time-consuming, creating prohibitive barriers for exploring new applications. This motivated the question:

RQ3: *How can we learn effective representations for human action recognition without extensive labeled data?*

In Paper II we propose a self-supervised representation learning method for pretraining action recognition models from unlabeled multi-view data, reducing the requirement for labeled data when tackling new downstream tasks.

While we have thus far focused exclusively on human pose estimation and analysis, human actions gain meaning through their environmental context: seeing someone reaching with their hand is not informative without knowing what objects are nearby. Moreover, many applications of multi-camera systems, from facility management to retail analytics, require detecting and localizing objects beyond humans. However, training a 3D object detector for every possible object category would be prohibitively expensive, requiring extensive data collection and annotation for each new deployment scenario. This motivates the need for open-vocabulary detection, where the system can detect arbitrary objects specified through natural language queries without retraining. Existing open-vocabulary 3D detection methods rely on dense point clouds from RGB-D sensors or hundreds of images, essentially building a complete 3D reconstruction of the scene before detection can occur. This is impractical for real-time systems with sparse cameras. This raises our fourth research question:

RQ4: *How can we achieve open-vocabulary 3D object detection from sparse RGB views?*

In Paper IV we develop a training-free approach for open-vocabulary 3D object detection from sparse RGB views that leverages pre-trained vision-language models and enforces multi-view consistency through optimization of monocular depth estimates. This enables the same camera system to provide comprehensive scene understanding, detecting both humans and arbitrary objects from just a few RGB views.

1.2 Thesis Outline

Chapter 2 describes the multi-camera system that motivated this research, detailing its architecture, applications, and the challenges that informed our research questions. Chapter 3 provides background on computer vision, multi-view geometry, deep learning, and human pose estimation necessary for understanding the technical contributions. Chapter 4 concludes with a synthesis of our contributions. The second part of the thesis contains the four papers that constitute the core technical contributions of this thesis.

Chapter 2

The MCS System

This thesis was conducted during the development of MCS (Multi-Camera System), a multi-view computer vision platform developed at Sony's R&D center in Lund.

The initial motivation for MCS arose from the need to explore new applications for human and animal activity recognition. Early experiments quickly showed that 2D analysis, including monocular 3D pose estimation, was insufficient for robust activity understanding in unconstrained environments. We concluded that the task would best be performed using multiple cameras capturing the subjects from different viewpoints. However, existing multi-view solutions were either too expensive, too complex to deploy, or too rigid to adapt to diverse environmental constraints. We needed a versatile, cost-effective platform that could be deployed quickly, reconfigured easily, and could be adapted to diverse use cases.

2.1 Architecture Overview

Figure 2.1 presents an overview of MCS' architecture. Distributed cameras capture synchronized video streams, and 2D human poses are estimated for each frame. During installation, multi-view pose sequences are used to calibrate the cameras. The person detections are matched across the camera views and 3D poses are reconstructed, for example via triangulation (Sec. 3.1.5). The 3D poses can be further processed, e.g. for activity recognition, or rendered for visualization.

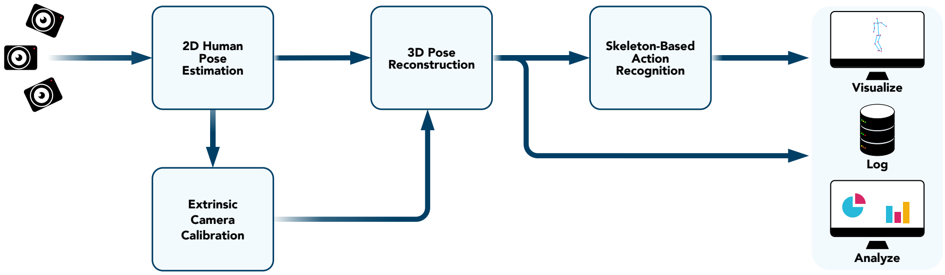


Figure 2.1: High-level overview of the MCS system architecture. Distributed cameras capture synchronized video streams, from which 2D human poses are estimated in each frame. During installation, sequences of multi-view poses are used for automatic extrinsic camera calibration. Once cameras are calibrated, the system matches detections across views and reconstructs 3D pose sequences, which can then be used for downstream tasks such as skeleton-based action recognition, visualization, and logging.

The MCS system has a modular architecture enabling its components, such as 2D pose estimation, 3D reconstruction and activity recognition, to be deployed in different topologies, as shown in Fig. 2.2. In the *centralized* configuration, cameras are connected to a single computer, enabling low-latency processing for simple room setups. In some environments, cables can not be easily run to the main computer, for example in large spaces, or in settings where cables should not be visible. In such situations, the system can be installed in a *decentralized* configuration, where each camera is connected to an edge processing unit performing 2D pose estimation. The 2D poses are streamed wirelessly to a main computer for 3D pose reconstruction and further analysis. In this configuration, only semantic data is exchanged over the network, and the images do not need to leave the edge computing units; this significantly reduces bandwidth requirements and can be seen as a privacy-preserving feature, as the 2D pose data does not contain any identifiable personal details. Finally, in very large spaces the system can be installed in a semi-distributed, *cascaded* configuration employing multiple computers, each handling 2D detection for several cameras while sending results to a central 3D reconstruction unit.

2.2 Applications

MCS' versatility has enabled deployment across very diverse application domains, each demonstrating different aspects of the system's capabilities and posing new technical challenges that informed the research presented in this thesis.

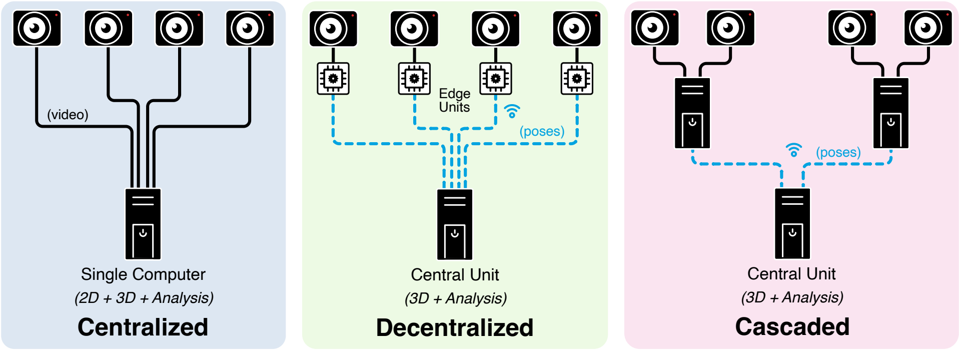


Figure 2.2: MCS network topologies for different deployment scenarios. Left: centralized configuration, where all cameras are connected to a single computer performing 2D detection, 3D reconstruction, and analysis with minimal latency. Middle: decentralized configuration, in which edge units perform 2D pose estimation locally and stream only 2D poses to a central unit for 3D reconstruction, reducing bandwidth requirements. Right: cascaded configuration for large spaces, combining several intermediate computers handling subsets of cameras before aggregating results in a central unit.

Realtime gym exercise recognition. The first major application of MCS was a proof-of-concept for Sony's Advagym service, providing automated exercise tracking and repetition counting in gym environments. The system was trained to recognize 15 different exercises, such as push-ups or squats, offering real-time feedback and session logging for multiple simultaneous users. The distributed architecture proved particularly valuable in gym settings, where extensive cabling is often impractical. This application demonstrates the system's ability to handle complex multi-person scenarios with frequent occlusions from equipment and other users. It also informed many design choices for the system, and in particular made us reflect on how the system would scale in the future: if it became a commercial application, how could camera systems be installed and calibrated in hundreds of gyms while keeping the costs manageable? How could we handle gathering data and training action recognition models to recognize the hundreds of exercises that are commonly used by gym-goers across the world? How could 3D reconstruction be implemented to be robust to the many occluding objects typically present in a gym?

Livestock research. Multiple MCS systems have been deployed at the Swedish University of Agricultural Sciences' research farm in Uppsala to study dairy cow welfare, e.g. for detecting brush use or measuring abnormal posture transitions [61, 74, 75]. This domain presents unique challenges beside detecting and tracking animals rather than humans, including harsh environmental conditions, and the need for continuous long-term remote operation with minimal maintenance. In particular, sending technicians to remote farms for recalibration would be economically infeasible.

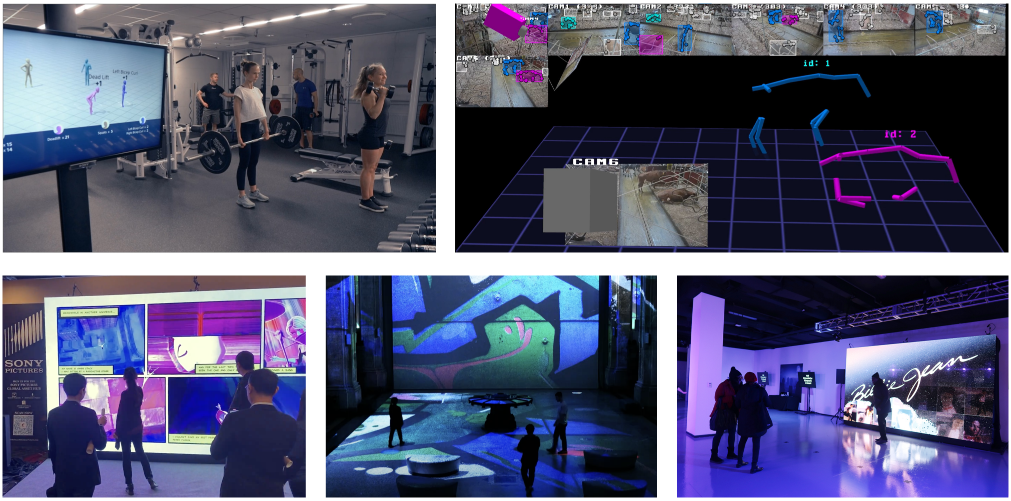


Figure 2.3: Example applications of MCS. Top-left: Real-time gym exercise recognition and counting. **Top-right:** Dairy cow 3D pose estimation and analysis. **Bottom:** Interactive entertainment experiences, including CinemaCon 2022, Hall des Lumières (New-York) and the Michael Jackson Thriller 40 Immersive Experience, where visitors' movements are captured and drive visual effects in real time.

Interactive experiences. MCS has powered several high-profile interactive installations, for example at the Hall des Lumières, and travelling promotional events for Sony entertainment properties, e.g. at CinemaCon 2022, the 2023 Licensing Expo, or the travelling Michael Jackson Thriller 40 Immersive Experience. These applications feature large screens surrounded by cameras that detect the visitors' movements to trigger interactive content. For example, performing Spider-Man's signature hand gesture would activate web effects on screen. These installations require robust performance in challenging environments with variable lighting, crowded conditions, and complex backgrounds. The interactivity of the applications also puts hard requirements on the latency of the whole system, from image capture through action recognition to rendering. The touring nature of these applications has been a particular challenge, requiring rapid deployment capabilities as the system must be installed, calibrated, and operational within a short time in completely new venues.

Each application has presented its own set of challenges and influenced the research agenda of this thesis.

2.3 From Applications to Research Questions

Across all applications, camera calibration emerged as a critical bottleneck. Traditional checkerboard calibration required specialized equipment that needed to be shipped to each location and trained operators who understood the calibration process. For the touring marketing installations, it was vital to reduce the complexity and time requirements for the calibration process. A major success factor for the MCS project was the ease of collaboration with teams of domain experts in remote locations and with different backgrounds. Having to start each partnership by a crash course on camera calibration or to personally assist with setup and maintenance of the systems would have been cumbersome and costly, and may have ultimately derailed the project. This directly motivated RQ1 and Paper I's research into human-pose-based calibration. By using natural human motion as the calibration target, we eliminated equipment requirements and reduced calibration to simply asking someone to walk around the space.

As we started to deploy MCS outside our lab, we soon realized that real-world environments violated many assumptions of traditional multi-view algorithms. Occlusions were the norm, not the exception, camera placement was dictated by practical constraints, lighting could not be easily adapted to our requirements, and background clutter confused 2D pose estimation models. With noisy or missing 2D joint detections, reconstructed 3D poses were corrupted, causing further analysis modules to fail. For example in the gym application, exercises were misclassified or some repetitions were missed, and spurious events would be triggered in interactive applications. These challenges inspired RQ2. The learning-based approach to 3D reconstruction proposed in Paper III, which explicitly handles missing 2D detections proved more robust than triangulation-based alternatives in challenging environments.

Using the reconstructed 3D poses for activity recognition involves training an activity classification model on large datasets of labelled pose sequences. When developing the gym application, we had to collect and manually annotate a large amount of examples for each new activity. Although we only supported 15 exercises in this prototype, this took a considerable amount of time and resources. Should the prototype become a commercial application, we realized, it would need to support many more exercises. This way of working would hardly scale. Simultaneously, as the system gained popularity within the company and we came into contact with new teams, many new ideas were generated. But by definition, when exploring a new domain few or no annotated data are available, which means that, even for exploring potential ideas, a considerable amount of time would have to be spent in gathering, curating and annotating data, before the idea could be evaluated. The annotation bottleneck severely limited our ability to explore new applications. This need to support rapid idea evaluation, and to be able to scale in case of success, triggered RQ3. While pondering it, we realized that our various installations continuously gathered data, although we did not have the time or the resources to curate them. By pre-training on unlabeled multi-view pose data, the method presented in Paper II makes it possible to obtain activity recognition models that can be rapidly fine-tuned to recognize new activities with minimal annotation effort, considerably reducing development time.

As the system matured and was now installed in various locations, a recurrent question was whether it could provide information about more than just the people moving in the scene. Providing data about the contents of the scenes, what objects were present and how they were used, could pave the way for many new applications. Efforts had been made to implement detectors for certain objects, but the work required to gather and annotate data for these object detectors meant that this fully-supervised approach would not help us realize such scene understanding. This led us to RQ4. Paper IV proposes a training-free pipeline for open-vocabulary 3D object detection using only sparse RGB images as input, a simple approach that is well-suited for continuous scene understanding under practical constraints.

2.4 Conclusion

The MCS system serves as both a practical platform for real-world deployment and a research vehicle for addressing fundamental challenges in multi-view computer vision. The diverse applications, from fitness tracking to interactive entertainment, provided a rich source of practical challenges that informed our research agenda.

Each deployment revealed limitations of existing approaches when confronted with real-world constraints. These experiences motivated the research contributions of this thesis, each addressing a specific pain point encountered during deployment.

This tight coupling between system development and research exemplifies how practical challenges can drive advances in computer vision, while ensuring that research contributions address real needs.

Chapter 3

Background

3.1 Computer Vision

This section introduces key concepts from computer vision and multi-view geometry used throughout this thesis. We cover the pinhole camera model and image formation, feature detection and matching between images, epipolar geometry for two-view systems, robust estimation using RANSAC, triangulation for 3D point reconstruction, and bundle adjustment for joint optimization of camera poses and scene structure.

These foundations are essential for understanding the multi-camera calibration method in Paper I, the multi-view 3D pose estimation approach in Paper III, and the 3D object detection system in Paper IV.

3.1.1 The Pinhole Camera Model

During image formation in a camera, the light reflected or emitted by visible 3D points is projected onto the image plane. While several mathematical models can describe the relation between the world coordinates of the 3D points and the 2D coordinates of their projections, the most widely used in computer vision is the *pinhole camera model*, a simple model that can, with some adaptations, be used to accurately describe the viewing geometry of most cameras. It describes how a perspective camera maps 3D points in the world to 2D points in the image. A key characteristic of perspective cameras is that they preserve straight lines. This simple model approximates a camera as a box with an arbitrarily small

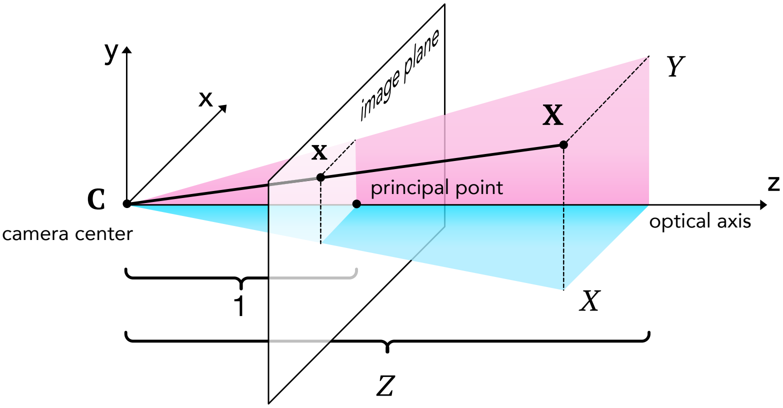


Figure 3.1: Pinhole camera model.

hole. Light rays passing through the hole, or camera center, project an image of the scene onto a back screen called the focal plane. This is the principle of the *camera obscura*, which has been used by artists and scientists since the 16th century to reproduce correct perspective. The line perpendicular to the focal plane and passing through the camera center is known as the optical axis and the point at which it intersects the image plane is called the principal point. The camera is represented by a 3D frame with its origin in the camera center, with the z-axis pointing forwards along the optical axis.

While the physical model projects the scene upside down, it is often more convenient to consider a virtual image plane that is parallel to the focal plane and located in front of the camera center, for example the normalized image plane at $z = 1$. As seen in Fig. 3.1, by similar triangles, the ray from the origin $\mathbf{C} = [0, 0, 0]^T$ to $\mathbf{X} = [X, Y, Z]^T$ intersects the image plane at $\mathbf{x} = [X/Z, Y/Z, 1]^T$.

In computer vision, it is common to represent points in Euclidean space $\mathbb{R}^n$ using their *homogeneous coordinates* in projective space $\mathbb{P}^n$, i.e. $\mathbf{x} = [x, y]^T \in \mathbb{R}^2$ can be represented by $\tilde{\mathbf{x}} = [x, y, 1]^T \in \mathbb{P}^2$ and $\mathbf{X} = [X, Y, Z]^T \in \mathbb{R}^3$ can be represented by $\tilde{\mathbf{X}} = [X, Y, Z, 1]^T \in \mathbb{P}^3$. Homogeneous coordinates enable to represent perspective projections as linear transformations, which greatly simplifies many geometric computations.

Using homogeneous coordinates, the mapping from 3D point coordinates to points on the image plane can be written in matrix form:

$$\tilde{\mathbf{x}} = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \frac{1}{Z} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}. \tag{3.1}$$

Converting from normalized image coordinates to pixel coordinates can be achieved by an affine transformation represented by the *camera calibration matrix*:

$$\tilde{\mathbf{u}} = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tilde{\mathbf{x}} = \mathbf{K}\tilde{\mathbf{x}}. \tag{3.2}$$

Here $\mathbf{K}$ contains the *intrinsic parameters* of the camera. f_x and f_y are the focal lengths in pixels, (c_x, c_y) is the principal point in pixel coordinates, which in ideal conditions would be in the centre of the image, and s is a skew factor that is often omitted in modern cameras.

Thus the transformation from 3D coordinates to pixel coordinates can be expressed

$$\mathbf{u} = \pi(\mathbf{K}\begin{bmatrix}\mathbf{I} & 0\end{bmatrix}\tilde{\mathbf{X}}) = \pi(\mathbf{P}\tilde{\mathbf{X}}), \tag{3.3}$$

where $\mathbf{P}$ is the *camera matrix*, and π is the perspective projection operator: $\pi(\begin{bmatrix}x, y, z\end{bmatrix}^T) = \begin{bmatrix}x/z, y/z\end{bmatrix}^T$.

Until now, we have considered a camera centered on the origin of the coordinate system and with viewing direction along the z-axis. For cameras with general position and orientation, the full camera matrix taking 3D point coordinates to pixel coordinates is given by

$$\mathbf{P} = \mathbf{K}\begin{bmatrix}R & \mathbf{t}\end{bmatrix}, \tag{3.4}$$

where $\mathbf{t} \in \mathbb{R}^3$ represents the camera translation and R is a 3×3 rotation matrix. R and $\mathbf{t}$ are called the *extrinsic parameters* of the camera.

While the pinhole camera model provides an excellent approximation for many applications, it is only a convenient abstraction, as the light passing through an arbitrarily small hole would not suffice to form an image in realistic settings. Actual cameras use larger openings to allow more light to enter, and lenses to concentrate and focus the light rays. The spherical shape of the lenses causes the light to be refracted differently at the center and at the edges. This effect, called *radial distortion*, makes straight lines appear curved in images. Tangential distortion, resulting from imperfect lens alignment during manufacturing, causes additional displacement perpendicular to the radial direction; however, in most modern cameras tangential distortion is often negligible compared to radial distortion.

In this thesis, we consider calibrated cameras for which the distortion parameters have been estimated and corrected for, effectively enabling the use of the simple pinhole camera model.

3.1.2 Feature Detection and Correspondence

Many problems in multi-view computer vision require finding and matching points of interest across images, for example for panorama stitching, camera pose estimation or sparse 3D reconstruction. The first step involves detecting candidate salient points or *keypoints*, i.e. points with well-defined position in the image and where the local structure of the image is rich in local information contents, typically at corners or regions with distinctive texture patterns. For each keypoint, a *feature descriptor* is computed, usually a vector calculated from the appearance of the image patch centered on the keypoint. Keypoints can then be matched between views by comparing their feature descriptors. For example, two keypoints can be matched if the Euclidean distance of their descriptors is below a given threshold.

A straightforward feature descriptor would be to take the raw pixel patches around the keypoints. However, small transformations can drastically affect feature distances. Ideally, features should be repeatable and invariant to perspective effects and illumination, so that different projections of the same 3D point can yield similar feature descriptors across viewpoints. Many algorithms for feature detection and description have been developed, such as SIFT [89], SURF [9], or ORB [120]. In particular, SIFT revolutionized feature detection and matching due to its invariance and robustness to changes in scale, rotation, and illumination, enabling the development of techniques for large-scale Structure from Motion (SfM), i.e. reconstruction of 3D scenes from 2D images.

Classical feature detectors are designed to detect salient points where image appearance shows high variation and do not typically find keypoints in textureless areas. Moreover, when two cameras have a wide baseline and observe the same object from very different viewpoints, the projections of the same 3D points in the images may have very different appearances. Recently, deep learning methods that learn to perform detection and description simultaneously have made substantial progress towards handling these issues [30, 35, 132, 165]. Although these methods can handle large changes in viewpoint, this is sometimes insuffi-

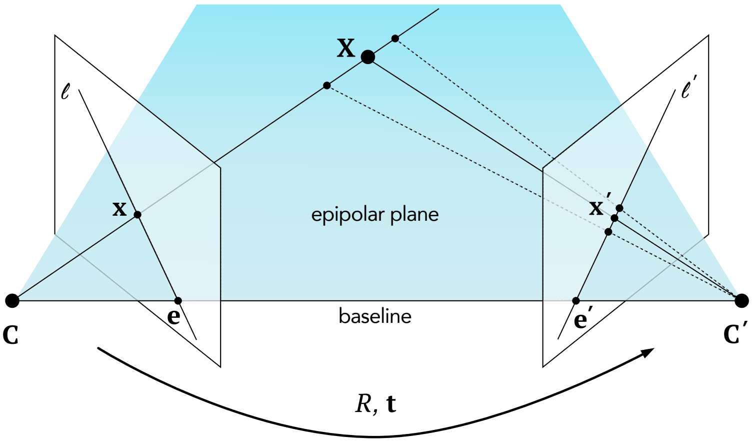


Figure 3.2: Epipolar geometry. If the projection of $\mathbf{X}$ on the left image is $\mathbf{x}$, then its projection on the right image must be located on the corresponding epipolar line ℓ'. The essential matrix allows to compute ℓ' from $\mathbf{x}$.

cient. In extreme cases, two cameras may see completely different sides of the same object, a case which no appearance-based descriptor can handle.

In Paper I, we use 2D body joint detections as keypoints, as their semantic nature has several advantages. Like in X-ray imaging, the same anatomical joint can be conceptually seen ”through” the body, making human joints more reliable for correspondence matching across wide baselines compared to traditional feature descriptors. Matching body keypoints across time frames is also straightforward, enabling the creation of tracks that can be used to enforce motion constraints. Finally, body keypoints also enable leveraging constraints on the structure of the body, such as constant limb lengths or plausible poses.

3.1.3 Epipolar Geometry

When two cameras observe the same 3D scene, the geometry relating the cameras, 3D scene points and the corresponding 2D observations is referred to as the *epipolar geometry* of the camera pair. This geometric relationship provides fundamental constraints that are essential for multi-view computer vision applications.

As illustrated in Fig. 3.2, consider two cameras observing the same 3D point $\mathbf{X}$, whose projection in each of the image planes is located at $\mathbf{x}$ and $\mathbf{x}'$ respectively. The camera centers are located at $\mathbf{C}$ and $\mathbf{C}'$, and the line between them is referred to as the *baseline*. The baseline intersects the image planes in two locations $\mathbf{e}$ and $\mathbf{e}'$ called the *epipoles* (i.e. the projection of the other camera's center on each image plane). The two camera centers and the point $\mathbf{X}$ span the *epipolar plane*, which intersects the image planes along the *epipolar lines* ℓ and ℓ'. As $\mathbf{x}$ and $\mathbf{x}'$ are necessarily on the epipolar plane, they are both located on the epipolar lines ℓ and ℓ', respectively.

Suppose that both the intrinsic and extrinsic parameters of the cameras are known, and that the coordinate system is that of the first camera. The projection matrices of the cameras are

$$\mathbf{P} = \mathbf{K}\begin{bmatrix} I & 0 \end{bmatrix}, \quad \mathbf{P}' = \mathbf{K}'\begin{bmatrix} R & \mathbf{t} \end{bmatrix}. \tag{3.5}$$

Given the homogeneous pixel coordinates $\tilde{\mathbf{u}}$, $\tilde{\mathbf{u}}'$ of observations in each image, the normalized image coordinates (or ray direction vectors) are

$$\tilde{\mathbf{x}} = \mathbf{K}^{-1}\tilde{\mathbf{u}} \sim \begin{bmatrix} I & 0 \end{bmatrix}\tilde{\mathbf{X}}, \quad \tilde{\mathbf{x}}' = \mathbf{K}'^{-1}\tilde{\mathbf{u}}' \sim \begin{bmatrix} R & \mathbf{t} \end{bmatrix}\tilde{\mathbf{X}}, \tag{3.6}$$

hence

$$\lambda\tilde{\mathbf{x}} = \begin{bmatrix} I & 0 \end{bmatrix}\tilde{\mathbf{X}} = \mathbf{X} \tag{3.7}$$

$$\lambda'\tilde{\mathbf{x}}' = \begin{bmatrix} R & \mathbf{t} \end{bmatrix}\tilde{\mathbf{X}} = R\mathbf{X} + \mathbf{t}, \tag{3.8}$$

where $\lambda, \lambda' \in \mathbb{R} \setminus \{0\}$ are unknown scalars as the position of $\mathbf{X}$ is unknown.

Substituting in the above equations, the observation $\tilde{\mathbf{x}}$ in the left image can be mapped onto the right image by the transformation

$$\lambda'\tilde{\mathbf{x}}' = \lambda R\tilde{\mathbf{x}} + \mathbf{t}. \tag{3.9}$$

Taking the cross-product with $\mathbf{t}$ on both sides gives

$$\lambda'\left[\mathbf{t}\right]_\times \tilde{\mathbf{x}}' = \left[\mathbf{t}\right]_\times (\lambda R\tilde{\mathbf{x}} + \mathbf{t}) = \lambda\left[\mathbf{t}\right]_\times R\tilde{\mathbf{x}}, \tag{3.10}$$

where $\left[\mathbf{t}\right]_\times$ is the skew symmetric matrix representing the cross-product with $\mathbf{t}$. Taking the dot product with $\tilde{\mathbf{x}}'$ on both sides gives a triple product with two identical elements on the left side, which is equal to 0, hence the nonnegative scalars λ and λ' can be eliminated:

$$\tilde{\mathbf{x}}'^T\left[\mathbf{t}\right]_\times \tilde{\mathbf{x}}' = \tilde{\mathbf{x}}'^T\left[\mathbf{t}\right]_\times R\tilde{\mathbf{x}} = \tilde{\mathbf{x}}'^T E\tilde{\mathbf{x}} = 0. \tag{3.11}$$

The matrix $E = \left[\mathbf{t}\right]_\times R$, called the *essential matrix*, encodes the relationships between corresponding points in the two images. Equation (3.11), known as the *epipolar constraint*, is powerful: even though the location of $\mathbf{X}$ is unknown, if we observe $\mathbf{x}$ in one image we know that potential matches for $\mathbf{x}$ must be on the corresponding epipolar line ℓ' in the other image, which restricts the search space from a 2D space to a 1D space. This dramatically reduces the computational complexity of feature matching.

The epipolar constraint also allows to calculate the epipolar lines: given $\tilde{\mathbf{x}}$, the associated epipolar line in the right image is $E^T\tilde{\mathbf{x}}$, and given $\tilde{\mathbf{x}}'$, the corresponding epipolar line in the left image is $E\tilde{\mathbf{x}}'$. The biased attention mechanism in Paper III encodes the epipolar constraint implicitly by giving higher weight to 2D observations lying closer to each other's epipolar line.

The essential matrix E deals with points expressed in normalized camera coordinates. Converting to pixel coordinates requires substituting the values of $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}'$ from Eq. (3.6), which yields

$$\tilde{\mathbf{u}}'^T K'^{-T} \left[\mathbf{t}\right]_\times RK^{-1}\tilde{\mathbf{u}} = \tilde{\mathbf{u}}'^T F\tilde{\mathbf{u}} = 0. \tag{3.12}$$

The matrix F is known as the *fundamental matrix*. Like the essential matrix, it represents the epipolar geometry of the cameras and allows to calculate the epipolar lines, but it does not require calibrating the cameras.

It is possible to retrieve E and F from point correspondences between two images using classical algorithms. Both matrices can be estimated up to scale using the 8-point algorithm [53, 88] with 8 point correspondences. Because of the cross-product matrix, E and F have rank 2. Leveraging this additional constraint by enforcing that $det(F) = 0$ enables to obtain F with only 7 point correspondences [52], which gives three possible solutions for F. The essential matrix has an additional *trace constraint* [29, 131] ensuring that it has two equal singular values

$$2EE^TE - \mathrm{tr}(EE^T)E = 0. \tag{3.13}$$

This allows to determine E with 5 correspondences for calibrated cameras using the 5-point algorithm [102], which gives 10 possible solutions.

Once the essential matrix E has been obtained, it can be decomposed into R and $\mathbf{t}$, up to a similarity transform. Four distinct solutions for R and $\mathbf{t}$ are possible, but only one enables to obtain points that are in front of both cameras by triangulation.

In Paper I, we estimate the essential matrix for a first pair of cameras with the 8-point algorithm using 2D body joint observations, and obtain their relative poses. This forms the foundation for our multi-camera calibration approach.

3.1.4 RANSAC

In real settings, detected keypoints will typically contain some noise and the correspondence set may contain mismatched keypoints that do not correspond to the same 3D object. These *outliers* would severely degrade the quality of the camera poses if they were used directly to estimate the camera matrices.

RANdom SAmple Consensus (RANSAC) [39] is an iterative method for robust model estimation that enables fitting a model to a set of observations while simultaneously filtering outliers. The key insight behind RANSAC is that while outliers can dramatically corrupt model estimation, a model fitted to a small subset containing only inliers should be supported by many of the remaining true inliers.

At each iteration, a minimal number of observations is sampled and used to estimate a model hypothesis. This hypothesis is then applied to all other observations and residuals are calculated. Observations with residuals smaller than a predefined threshold are considered inliers and constitute the *consensus set* of the hypothesis. After a given number of iterations the hypothesis that yielded the largest consensus set is selected as the solution. Finally, a refined solution can be estimated using the entire consensus set of the selected hypothesis.

The parameters of the RANSAC procedure are the number of sampled points s, the residual threshold τ and the number of iterations N. Sampling a minimal number of observations (e.g. 8 in the case of the 8-point algorithm) increases the probability of sampling only true inliers, which should yield a model with many true inliers in the consensus set. The inlier threshold τ is often set using prior domain knowledge about expected noise levels and can be difficult to set optimally in real-world applications. Some methods have been proposed to adaptively estimate τ during RANSAC [34, 59]. For a given inlier ratio ω, the number N of iterations that guarantees with probability p (e.g., $p = 0.99$) that at least one sample of s observations will contain only true inliers is $N = \left\lceil \frac{\log(1-p)}{\log(1-\omega^s)} \right\rceil$ [39]. Some variants of RANSAC are also able to estimate N adaptively based on the observed inlier ratios during the RANSAC process [54].

RANSAC is essential in multi-view geometry because feature matching algorithms inevitably produce false correspondences, especially in challenging scenarios with repetitive textures, lighting changes, or wide baselines. Without robust estimation, these outliers would make reliable camera pose estimation impossible.

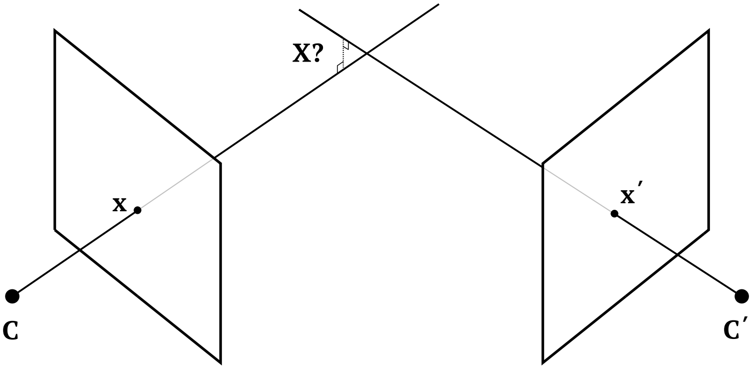


Figure 3.3: Triangulation. $\mathbf{X}$ should be found at the intersection of the rays passing through the camera centers and the 2D observations $\mathbf{x}$ and $\mathbf{x}'$. However in real settings the observations will often be noisy and the rays may not intersect.

3.1.5 Triangulation

Once camera poses are known and point correspondences have been established, the 3D coordinates of scene points can be recovered through *triangulation.*

Figure 3.3 illustrates the problem with two cameras. Intuitively, the solution should be found at the intersection of the rays passing through the camera centers and the 2D observations. In real settings, however, the two rays will typically not have a point of intersection, as the observations are usually noisy and the camera matrices may not be perfectly estimated.

The optimal solution for the triangulation problem, assuming Gaussian noise of the observations, minimizes the ℓ_2-norm of the reprojection error, and a simple solution exists for two views [55], but finding the global optimum for more than two views is an active subject of research [78].

While it does not guarantee an optimal solution, the *Direct Linear Transform* (DLT) method [54] is a widely-used solution to the triangulation problem as it is fast and generalizes well to multiple views.

Assume that we know the camera matrices P_i of $n \geq 2$ cameras and n 2D correspondences $\mathbf{u_i}$ for a 3D point $\mathbf{X}$, then

$$\tilde{\mathbf{u_i}} = \begin{bmatrix} u_i & v_i & 1 \end{bmatrix}^T \sim P_i \tilde{\mathbf{X}}. \tag{3.14}$$

The vectors are collinear, hence

$$\tilde{\mathbf{u_i}} \times P_i\tilde{\mathbf{X}} = \left[\tilde{\mathbf{u_i}}\right]_\times P_i\tilde{\mathbf{X}} = \begin{bmatrix} 0 & -1 & v_i \\ 1 & 0 & -u_i \\ -v_i & u_i & 0 \end{bmatrix} P_i\tilde{\mathbf{X}} = 0. \tag{3.15}$$

This gives three linear equations for each camera, of which two are independent:

$$u_i({\mathbf{p_i}}^{3T}\tilde{\mathbf{X}}) - ({\mathbf{p_i}}^{1T}\tilde{\mathbf{X}}) = 0 \tag{3.16}$$

$$v_i({\mathbf{p_i}}^{3T}\tilde{\mathbf{X}}) - ({\mathbf{p_i}}^{2T}\tilde{\mathbf{X}}) = 0 \tag{3.17}$$

$$u_i({\mathbf{p_i}}^{2T}\tilde{\mathbf{X}}) - v_i({\mathbf{p_i}}^{1T}\tilde{\mathbf{X}}) = 0. \tag{3.18}$$

Combining these equations gives a system of linear equations in the form

$$A\tilde{\mathbf{X}} = \begin{bmatrix} u_1{\mathbf{p}_1}^{3T} - {\mathbf{p}_1}^{1T} \\ v_1{\mathbf{p}_1}^{3T} - {\mathbf{p}_1}^{2T} \\ \cdots \\ u_n{\mathbf{p_n}}^{3T} - {\mathbf{p_n}}^{1T} \\ v_n{\mathbf{p_n}}^{3T} - {\mathbf{p_n}}^{2T} \end{bmatrix} \tilde{\mathbf{X}} = 0, \tag{3.19}$$

giving $2n$ equations for three unknown coordinates. In real applications the 2D observations will propably be noisy and there may not be an exact solution, so we will instead write $A\tilde{\mathbf{X}} = \mathbf{w}$ and solve for $\tilde{\mathbf{X}}$ such that the norm of $\mathbf{w}$ is minimized. This can be solved by determining the singular-value decomposition (SVD) of A and choosing $\tilde{\mathbf{X}}$ as the right singular vector corresponding to the smallest singular value of A.

Linear triangulation methods typically produce acceptable 3D estimates that can be used to initialize iterative non-linear refinement methods. In Paper I, the 3D positions of human body joints visible in a first camera pair are initialized using this method, and are optimized together with the relative poses of the cameras in a *bundle adjustment* stage (Sec. 3.1.6).

The results produced by this method degrade when the noise of the 2D observation increases. This is a problem in particular for 3D human pose reconstruction, as the accuracy of 2D pose detections depends on factors such as lighting conditions, occlusions, and viewing angle. Iskakov et al. alleviate the issue of noisy detections by weighting each row in Eq. (3.19) with the confidence score of the joints predicted by the detector [64]. However, as the number of views decreases, triangulation-based methods struggle, and when there are fewer than two views they simply cannot produce a result. In Paper III we therefore propose a learning-based method to solve the 3D pose reconstruction problem instead of relying on triangulation.

3.1.6 Bundle Adjustment

The methods we have seen so far can be used to sequentially estimate the relative poses of many cameras. After estimating the essential matrix for a camera pair, the relative poses of the pair can be used to triangulate the coordinates of 3D points that are visible in both cameras. For each subsequent camera, triangulated 3D points that are visible in that camera are matched to 2D observations in that image, and used to estimate the camera pose using algorithms such as EPnP [81]. The set of 3D points is expanded after the addition of each camera by triangulating more visible points.

However, this incremental approach only uses partial information at each stage, and errors due to noise and occlusions will typically accumulate. Moreover, the sequential estimation process does not enforce global consistency across all cameras and observations simultaneously.

Bundle adjustment is a nonlinear optimization method for producing a coherent optimized reconstruction using all the cameras and 3D points simultaneously. The term ”bundle” refers to the bundle of light rays connecting each 3D point to its projections in multiple camera views. The method minimizes the reprojection error across all cameras and 3D points:

$$\min_{P_i, \tilde{\mathbf{X}}_j} \sum_{i,j} ||\mathbf{u}_{ij} - \pi(P_i \tilde{\mathbf{X}}_j)||^2, \tag{3.20}$$

where P_i are the camera matrices, $\tilde{\mathbf{X}}_j$ are the 3D points in homogeneous coordinates, $\mathbf{u}_{ij}$ are the observed 2D projections, and π denotes the pinhole projection operator defined in Sec. 3.1.1.

This optimization problem can be solved using nonlinear least squares methods such as Levenberg-Marquardt [41, 93], which iteratively refine the camera poses and 3D point positions.

A similar process is used in Paper I to refine the initial estimates of camera poses given 2D body joint observations. In Paper I, we use gradient descent for bundle adjustment to enable leveraging domain-specific priors about observation noise, human motion and anatomy, which helps overcome the challenges posed by noisy joint detections.

3.2 Deep Learning

Deep learning has revolutionized computer vision since the breakthrough achieved by AlexNet in 2012 [73]. The superior performance of deep neural networks (DNNs) in detection, classification, and regression tasks has made them the dominant approach in modern computer vision pipelines. This section presents fundamental concepts and architectures used in this thesis.

The strength of deep neural networks lies in their ability to learn complex, non-linear mappings through the composition of multiple layers of simple operations. According to the Universal Approximation Theorem [60], neural networks with sufficient width can approximate any continuous function to arbitrary precision. In practice, depth rather than width has proven more effective, leading to the development of increasingly deep architectures.

The papers presented in this thesis use several neural network architectures: graph neural networks (GNNs) model skeletal structure for activity recognition in Paper II, Transformers enable robust multi-view 3D pose reconstruction in Paper III, while multilayer perceptrons (MLPs) and convolutional neural networks (CNNs) are the building blocks of all the architectures used here. The learning paradigms span supervised learning for pose estimation, self-supervised learning for action representation (Paper II) and semi-supervised approaches for leveraging limited labeled data.

3.2.1 Multilayer Perceptron

The Multilayer Perceptron (MLP) forms the foundation of modern deep learning architectures. An MLP consists of multiple layers of neurons, where each layer performs a linear transformation f_i followed by a non-linear activation function. Formally, a layer i in an MLP can be described as

$$\mathbf{x}_{i+1} = \sigma(f_i(\mathbf{x}_i, W_i, \mathbf{b}_i)) = \sigma(W_i\mathbf{x}_i + \mathbf{b}_i), \tag{3.21}$$

where $\mathbf{x}_i \in \mathbb{R}^{N_i}$ is the input vector, $W_i \in \mathbb{R}^{M_i \times N_i}$ are the learnable weights, $\mathbf{b}_i \in \mathbb{R}^{M_i}$ is the bias vector, and σ is a non-linear activation function. The dimensions satisfy $N_{i+1} = M_i$, ensuring compatibility between consecutive layers. The non-linearity introduced by the activation function is essential for the expressive power of neural networks. Indeed, without it even a very deep multi-layer network would collapse to just one linear layer and lose much of its representational power. A common choice is the *Rectified Linear Unit* (ReLU) $\sigma(\mathbf{x}) = \max(0, \mathbf{x})$, which provides computational efficiency and addresses van-

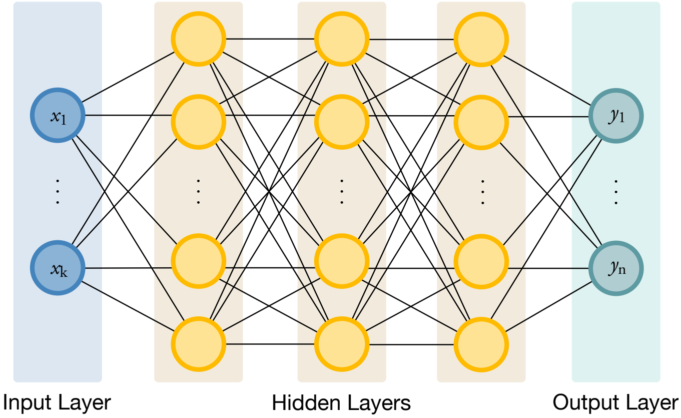


Figure 3.4: Multilayer Perceptron. Illustration of an MLP with one input layer, three hidden layers and one output layer.

ishing gradient problems. A complete MLP with D layers can be expressed as the composition

$$G(\mathbf{x}) = \sigma(f_D(\sigma(f_{D-1}(\dots\sigma(f_1(\mathbf{x}))\dots)))). \tag{3.22}$$

The depth D and layer widths M_i govern the model's capacity.

3.2.2 Regression and Classification

The output layer of the MLP produces the final prediction of the network and has the same number of neurons as the dimension of the output vector. In general, its activation function is different from those used in the hidden layers of the model. For regression tasks, the last layer typically has a linear activation function (i.e. no non-linearity). For classification tasks with C classes, the final layer uses the softmax activation function

$$\operatorname{softmax}(\mathbf{x}) = \frac{\left[e^{x_1}, \dots, e^{x_C}\right]^T}{\sum_{j=1}^{C} e^{x_j}}, \tag{3.23}$$

which produces a score vector that can be interpreted as a probability distribution over classes since each element is in the range $[0, 1]$ and the sum of the elements is 1.

MLPs serve as the final classification or regression layers in most architectures used throughout this thesis.

3.2.3 Convolutional Neural Networks

As fully connected layers treat inputs as flattened vectors, every output element interacts with input element. Many signals such as time series, images or 3D voxel grids are spatially structured; images, for example, often exhibit strong local correlations. The standard network architecture for such data are Convolutional neural networks (CNNs), which differ from fully connected networks by their *local connectivity* and *weight sharing.*

CNNs build on the discrete convolution operation, which can be thought of as a filter or *convolution kernel* sliding across the input; at each location the weighted average of the neighbourhood is computed based on the kernel weights, producing a response field. Convolution is performed with different kernels simultaneously, creating a multidmensional *feature map* that captures different aspects of the input. Different neurons along the depth dimension may activate in presence of various patterns, e.g. oriented edges or blobs of color.

Formally, for a two-dimensional input $\mathbf{x}_l \in \mathbb{R}^{H \times W \times C_l}$ with height H, width W, and C channels, a convolutional layer applies C_{l+1} kernels $K_k \in \mathbb{R}^{F \times F \times C_l}$. The convolution operation (which is technically a cross-correlation) at location (i, j) is defined as

$$(\mathbf{x}_{l+1} \star K_k)(i,j) = \sum_m \sum_n \sum_c \mathbf{x}_l(i+m, j+n, c) K_k(m,n,c), \tag{3.24}$$

and the complete action of the convolutional layer to transforms the input map $\mathbf{x}_l$ to the output map $\mathbf{x}_{l+1}$ is

$$\mathbf{x}_{l+1} = h_l(\mathbf{x}_l) = \sigma(\left[\mathbf{x} \star K_1, \ldots, \mathbf{x} \star K_{C_{l+1}}\right]^T + \mathbf{b}_l), \tag{3.25}$$

where σ is a non-linear activation function and $\mathbf{b}_l$ is a bias vector of dimension C_{l+1}, i.e. containing one shared scalar per channel of the output feature map.

Multiple convolutional layers can be stacked to form deep networks that learn hierarchical feature representations. Early layers detect simple patterns like edges and textures, while deeper layers combine these basic features into more complex visual concepts.

The kernel is typically much smaller than the input, so that at each location the output feature map depends only on a small neighbourhood. This contrasts with fully-connected (FC) layers that connect every neuron in one layer to every neuron in the previous layer. This *local connectivity* leads to lower computational complexity while preserving spatial coherence.

The same salient features producing a high response from a kernel at some location gives the same response for the same kernel at other locations. This *weight sharing* dramatically reduces the number of learnable parameters compared to fully-connected layers, making CNNs more data-efficient and less prone to overfitting.

A fundamental property of convolutional layers is *translation equivariance*: if the input is translated by some amount, the output feature map translates by a corresponding amount. This property allows CNNs trained on objects in different positions to generalize effectively. Translation equivariance is essential for computer vision applications such as object detection or segmentation, where the object's location within an image should not affect its detection.

Convolutional layers are ususally used together with other components that play an important role in neural networks.

Pooling. Pooling operations reduce the size of feature maps by using some function to summarize subregions, such as taking the average or the maximum value. *Max pooling* selects the maximum activation within each pooling window. *Average pooling* computes the mean instead of the maximum. Pooling makes the representation less sensitive to small spatial shifts, while also reducing computational load.

Batch Normalization. Batch normalization normalizes the activations of each layer within a mini-batch [62]. This addresses the problem of internal covariate shift, where a change in scaling in early layers would impact following layers, by enforcing a more consistent distribution of activations. For a mini-batch of inputs, it normalizes each feature dimension to have empirical mean zero and unit variance, then applies learnable scale and shift parameters:

$$\hat{\mathbf{x}} = \gamma \frac{\mathbf{x} - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta, \tag{3.26}$$

where μ_B and σ_B are the empirical mean and variance of the mini-batch, γ and β are learnable parameters, and ϵ is a small constant for numerical stability. Batch normalization accelerates training, enables higher learning rates, and reduces sensitivity to initialization. It plays an important part in the self-supervised learning method of Paper II, where it acts as an implicit contrast term [50].

3.2.4 Graph Convolutional Networks

Graph Convolutional Networks (GCNs) extend the concept of convolution from grids to non-Euclidean, graph-structured data such as social networks, molecules or, in our case, human skeletons. In a graph $G = (V, E)$ with nodes V and edges E, a graph convolution operation updates each node's features by aggregating features from its neighbours, as defined by the graph's adjacency, using shared learnable weights. There are two main paradigms for GCNs: *spectral* methods and *spatial* methods. The skeleton-based action recognition model used in Paper II [162] is based on the GCN layer defined by Kipf and Welling [69], a first-order approximation of spectral convolution on graphs that is computationally equivalent to a spatial *message-passing* layer. Intuitively, the GCN aggregates a node's neighbourhood information just as image convolution aggregates a pixel's local patch. Formally, the transformation of the nodes' features at each layer can be written in matrix form as

$$H_{l+1} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H_l W_l\right), \tag{3.27}$$

where $\tilde{A} = A + I$ is the graph's adjacency matrix A with added self-connections (identity I), $\tilde{D}$ is the degree matrix of $\tilde{A}$, H_l is the matrix of node features at layer l and W_l is the learnable weight matrix. This formula shows that each node's new features $h_{l+1}^{(i)}$ are a weighted sum of its own features $h_l^{(i)}$ and its neighbours' features $h_l^{(j)}$, normalized by the neighbour counts.

3.2.5 Transformers

The Transformer [146] has proven to be a powerful architecture with broad applications in various fields, from natural language processing [36, 146] to computer vision [17, 33].

A Transformer acts on sets of *tokens*, vectors that represent data points such as word embeddings, image patches or 3D points. The core innovation underlying Transformers is the *attention* mechanism, which enables tokens to *'attend'* to each other. It computes a weighted sum of a set of tokens based on their relevance to a particular query. In particular, scaled dot-product attention calculates the relevance score of each key $\mathbf{K}$ with respect to a query $\mathbf{Q}$:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{3.28}$$

where queries $\mathbf{Q}$, keys $\mathbf{K}$ and values $\mathbf{V}$ are projections of the input tokens and d_k is the dimension of the feature vectors.

In the *self-attention* mechanism $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ are three representations from the same set of input tokens: $\mathbf{Q} = W_Q\mathbf{X}^T$, $\mathbf{K} = W_K\mathbf{X}^T$, and $\mathbf{V} = W_V\mathbf{X}^T$, where $W_Q, W_K, W_V \in \mathbb{R}^{d\times d_k}$ are learned projection matrices and $\mathbf{X} \in \mathbb{R}^{N\times d}$. Thus all input tokens attend to each other. With *cross-attention*, two distinct sets of tokens interact with each other, and in general $\mathbf{Q} = W_Q\mathbf{Y}^T$, $\mathbf{K} = W_K\mathbf{X}^T$ and $\mathbf{V} = W_V\mathbf{X}^T$, where $W_Q \in \mathbb{R}^{d_2\times d_k}$, $W_K, W_V \in \mathbb{R}^{d_1\times d_k}$, $\mathbf{X} \in \mathbb{R}^{N\times d_1}$ and $\mathbf{Y} \in \mathbb{R}^{N\times d_2}$, i.e. the input tokens $\mathbf{X}$ are queried using tokens $\mathbf{Y}$.

Rather than using a single attention function, Transformers usually employ multiple attention heads that can focus on different types of relationships

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, , \text{head}_h)W^O, \tag{3.29}$$

where h is the number of heads, each $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$ and W^O is a final linear projection. The W_i^Q, W_i^K and W_i^V project the queries, keys and values h times and the attention function is performed in parallel on these projections.

The attention mechanism makes the attention layer permutation-equivariant. However, in many applications the order or relative position of the input is important, for example the order of the words in a sentence in natural language processing, or the location of patches in an image for vision tasks. It is therefore common practice to attach to the tokens a *positional encoding* indicating their relative positions.

In Paper III, we adapt Transformers for robust multi-view 3D pose estimation. Each camera view provides a sequence of 2D joint detections, and the Transformer learns to aggregate information across views while handling missing or corrupted observations. The Transformer encoder fuses multi-view and temporal information through self-attention, while a decoder queries this representation to predict 3D joint positions. We introduce geometry-biased attention that incorporates the geometric relationships between camera views.

3.2.6 Supervised Learning

The dominant approach for training deep neural networks is *supervised learning.* Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ of pairs of input data points $\mathbf{x}_i$ and corresponding *ground truth* annotations $\mathbf{y}_i$, the objective is to find a function $f(\mathbf{x}; \theta)$, parameterized by θ, that can accurately predict the target $\mathbf{y}$ for a new, unseen input $\mathbf{x}$. This learning problem can be formulated as an optimization problem where we seek to find the parameters $\theta^\star$ that minimize the expected discrepancy between output and ground truth, or *loss*, over the training dataset:

$$\theta^\star = \arg\min_\theta \sum_{i=1}^{N} \mathcal{L}(f(x_i;\theta), y_i), \tag{3.30}$$

where the choice of loss function $\mathcal{L}$ depends on the learning task. For regression problems, it is common to use the mean squared error (MSE) loss

$$\mathcal{L}_{\text{MSE}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \underbrace{||y_i - f(x_i;\theta)||^2}_{\mathcal{L}_{\text{MSE}}(f(x_i;\theta),y_i)}, \tag{3.31}$$

while for classification tasks the cross-entropy loss provides a probabilistically principled objective

$$\mathcal{L}_{\text{CE}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \underbrace{\sum_{c=1}^{C} y_{i,c}\log(p_{i,c})}_{\mathcal{L}_{\text{CE}}(f(x_i;\theta),y_i)}, \tag{3.32}$$

where $y_{i,c}$ is the one-hot encoded ground truth and $p_{i,c}$ is the likelihood for class c predicted by the softmax layer.

The minimization problem is generally non-convex and high-dimensional, and will often result in a local optimum or a saddle point rather than a globally optimal set of parameters. The standard approach for finding $\theta^\star$ combines gradient-based optimization with the *back-propagation* algorithm [123] to compute gradients efficiently through the computational graph. The most popular optimization methods are based on *gradient descent* which, starting from a randomly initialized set of parameters θ_0, iteratively updates the parameters in the direction that decreases the loss most rapidly. Since the gradient $\nabla_\theta \mathcal{L}(\theta)$ indicates the direction of steepest increase of the loss function, the parameters are updated in the opposite direction. The parameter update for each iteration of gradient descent on the complete training dataset can be written

$$\theta_{t+1} = \theta_t - \frac{1}{N}\sum_{i=1}^{N} \nabla_\theta \mathcal{L}(x_i, y_i; \theta_t), \tag{3.33}$$

where t indexes the training iteration and $\eta > 0$ is the learning rate that controls the step size. The learning rate requires careful tuning: values that are too large may cause the optimization to overshoot minima or become unstable, while values that are too small may result in prohibitively slow convergence. Many training methods use learning rate schedules that adapt η during training, starting with a large value for rapid initial progress and decreasing it during training to ensure convergence.

Computing gradients over the entire dataset becomes intractable for large datasets. *Stochastic Gradient Descent* (SGD) addresses this by approximating the full gradient using randomly sampled *mini-batches*, or even a single data point, at each iteration. For a mini-batch of size $m \ll N$, the update step becomes

$$\theta_{t+1} = \theta_t - \eta \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \mathcal{L}(x_i, y_i; \theta_t). \tag{3.34}$$

While individual mini-batch gradients are noisy estimates of the full gradient, this stochasticity can actually help escape local minima and often leads to better generalization. Many variations of the SGD algorithm have been proposed. *Momentum* smoothes the noisy gradient estimates by keeping a running estimate, accelerating SGD in the relevant direction [123]. Adam (Adaptive Moment Estimation) uses exponential averages of the first and second moments of the gradient to adapt the learning rates of different parameters of the network [68]. Its robustness and minimal tuning requirements make it a popular optimization method for deep learning.

Proper evaluation requires careful data partitioning into training, validation, and test sets. The *training set* is used for parameter optimization, the *validation set* for hyperparameter tuning and early stopping, and the *test set* for final performance evaluation. This separation is important to avoid overfitting and ensure reliable assessment of generalization performance.

The fundamental challenge in supervised learning is achieving good generalization to unseen data. Networks with sufficient capacity can memorize the training data, leading to poor performance on new examples. Regularization techniques such as weight decay, dropout and data augmentation help combat overfitting by constraining model complexity or introducing variation during training.

A significant practical limitation of supervised learning is its dependence on manually annotated datasets, which represents a labor-intensive, time-consuming, and expensive process that often becomes the bottleneck in developing new applications. This limitation motivated us to study self-supervised learning in Paper II.

3.2.7 Self-Supervised Learning

Self-supervised learning aims at learning transferable representations by extracting supervisory signals from the data itself rather than external annotations. The fundamental principle of self-supervised learning is to construct learning tasks where the ground truth can be automatically derived from the input data itself, eliminating the need for manual annotation while still providing meaningful learning objectives. These pretext tasks are designed to encourage the model to learn representations that capture important structural properties of the data domain, such as spatial relationships, temporal dynamics, or semantic consistency.

Early methods used pretext tasks related to high-level image understanding, for example reconstructing the original data by denoising [147], inpainting [105], colorizing [77, 166, 167] or solving jigsaw puzzles [31, 103], or predicting transformations [43].

Contrastive learning methods [19, 20, 23, 57, 99, 144] represent one of the most successful approaches to self-supervised learning, based on the principle of learning representations that bring similar examples closer together while pushing dissimilar examples apart in the representation space. The key insight is that similar examples (positive pairs) should have similar representations, while dissimilar examples (negative pairs) should have dissimilar representations. Positive pairs are constructed by augmenting the same input in two different ways, while two different data points in the dataset are assumed to be dissimilar. Given a batch of N inputs, each item is augmented in two different ways to construct a positive pair. For each positive pair of points, all remaining $2(N-1)$ points are considered negative examples. For a positive pair $(\mathbf{z}_i, \mathbf{z}_j)$, the contrastive loss can be written

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)} \,, \tag{3.35}$$

where $\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k) = \frac{\boldsymbol{z}_i^T \cdot \boldsymbol{z}_k}{\|\boldsymbol{z}_i\| \cdot \|\boldsymbol{z}_k\|}$ is the cosine similarity between vectors $\boldsymbol{z}_i$ and $\boldsymbol{z}_k$, and τ is a temperature parameter.

Contrastive learning methods face several practical challenges, particularly the computational overhead of processing large numbers of negative samples. SimCLR [19, 20] uses large batch sizes, and MoCo [23, 57] keeps a large queue of past representations as negative samples. Several recent works showed that useful representations can be learnt without the need for explicit negative samples [18, 22, 50]. Bootstrap Your Own Latent (BYOL) [50], for example, uses a momentum-updated target network that provides stable learning targets for an online network to predict.

Paper II demonstrates how BYOL can be adapted for self-supervised learning of action representations from 3D pose sequences. We designed data augmentation strategies that preserve the semantic content of actions while providing sufficient variation to drive representation learning. Geometric augmentations, such as rotation and scaling, preserve the essential structure of human motion while creating diverse training examples. Temporal augmentations, such as subsampling or temporal jittering, encourage the model to learn representations that are robust to timing variations while maintaining sensitivity to the essential dynamics of different actions. The multi-view context provides particularly rich opportunities for self-supervised learning, since geometric consistency across views provides a natural source of supervision, as observations of the same 3D pose from different camera viewpoints should yield semantically consistent features.

3.2.8 Foundation Models and Vision-Language Understanding

The emergence of foundation models, trained on massive datasets and capable of adapting to diverse downstream tasks, has fundamentally changed the landscape of machine learning and computer vision. These models demonstrate that large-scale pretraining on diverse data can produce representations that transfer effectively to a wide range of applications, often with minimal or no task-specific fine-tuning [10].

Vision-language models such as CLIP (Contrastive Language-Image Pre-training) [112] represent a particularly important class of foundation models that learn joint representations of images and text through contrastive learning on ”internet-scale” data (several hundred million images and associated captions). CLIP demonstrates remarkable zero-shot capabilities, and is able to classify images into arbitrary categories specified through natural language descriptions without task-specific training.

CLIP uses separate encoders for images and text that are trained to map inputs to a shared embedding space, where semantically related content has high similarity. The training objective encourages high similarity between images and their associated captions while discouraging similarity between images and unrelated text. This simple but powerful approach enables the model to learn rich visual representations that are grounded in human language and can be queried using natural language descriptions.

In contrast to image classifiers trained in a supervised manner on labeled images, CLIP is trained by *weak supervision*. Depending on the context in which the captions describing the images were produced, they may focus on different aspects of the images, describe som parts in more detail or disregard some objects. Thanks to the immense scale of the dataset, the model sees the same concept in various settings in many images with different descriptions, and is able to capture rich semantic information without task-specific supervision.

Segment Anything is a foundation model for image segmentation that was trained to segment any object given prompts [72] (c.f. Sec. 3.3.2). It used a model-in-the-loop strategy to create a large dataset with a data engine combining human annotation, semi-automatic annotation and fully automatic annotation. This data engine strategy was successfully used in other domains, such as monocular depth estimation [153, 163].

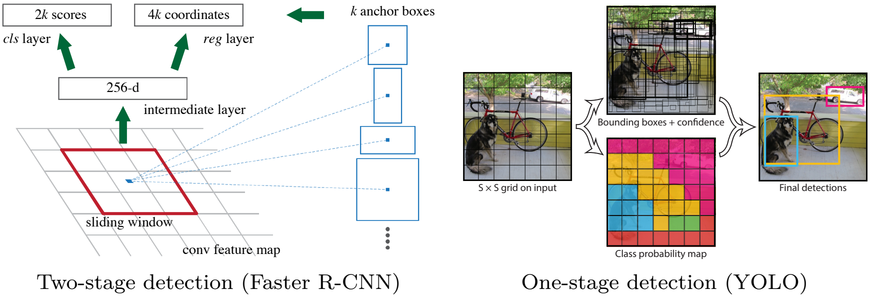
Paper IV demonstrates how foundation models can enable open-vocabulary 3D object detection from sparse multi-view RGB images without task-specific training. It leverages foundation models for 2D object detection from textual descriptions, object segmentation and monocular depth estimation. During depth refinement, the rich semantic embeddings learned by CLIP are used to guide the optimization, establishing cross-view correspondences even in cases where the cameras see different sides of objects.

3.3 Object Detection, Segmentation and 3D Scene Understanding

3.3.1 2D Object Detection

Object detection addresses the task of identifying and localizing objects in observed scenes. This involves predicting the semantic class of the objects along with the *bounding boxes* surrounding them. For 2D detection, bounding boxes are often described by their center coordinates, width and height.

Early methods performed object detection by extracting and classifying handcrafted features from images [27, 148]. Deep-learning-based 2D object detection methods generally follow one of two architectural paradigms. Two-stage detection methods, exemplified by the R-CNN family [44–46, 56, 115], decompose the problem into region proposal generation followed by classification and localization refinement. Faster R-CNN introduced the Region Proposal Network (RPN) that slides over extracted feature maps and predict region proposals based on



Two-stage detection (Faster R-CNN)

One-stage detection (YOLO)

Figure 3.5: Left: Two-stage object detection methods predict region proposals then pool features to predict the class and refine box parameters of each proposal. **Right:** One-stage methods predict bounding box coordinates and classes simultaneously. Images from [115] and [113].

predefined anchors at each location [115]. In a second stage, features inside each region of interest are pooled from the feature map and used to predict object class and refine the bounding box parameters.

Single-stage methods such as YOLO [113] or SSD [85] take a different approach by directly predicting a fixed number of object descriptors for every spatial location in the downsampled feature representation. These methods use an objectness probability to determine whether each location actually contains an object.

Both paradigms typically employ *non-maximum suppression* to handle multiple overlapping detections of the same object by selecting the detection with the highest confidence score among highly overlapping candidates.

Recent developments have introduced Transformer-based detection methods like DETR [17], which reformulate object detection as a set prediction problem using attention mechanisms to directly predict object locations without requiring hand-crafted components like anchor generation or non-maximum suppression.

Traditional object detectors require training on predefined object classes, limiting their applicability to new domains or novel object categories. Open-vocabulary 2D object detection has recently emerged as a powerful paradigm to overcome this limitation, enabling the localization of arbitrary object classes using natural language queries. Early methods such as ViLD [51] and RegionCLIP [171] demonstrated that distilling knowledge from pretrained vision-language models like CLIP [112] could transfer zero-shot recognition capabilities to standard detectors, while Detic [172] showed that large-scale image-level supervision could greatly expand the detectable vocabulary. More recent works, including OWL-ViT [97] and its successor OWLv2 [98], proposed single-stage transformer detectors built on CLIP-like backbones, offering strong zero-shot transfer and scalability to billions of image-text pairs. In parallel, Grounding DINO [84] buildt upon the DETR framework [17] by integrating natural language as a conditioning signal via cross-modal attention, and YOLO-World [24] augmented the YOLOv8 detector with a vision-language branch and a region-text contrastive loss, demonstrating zero-shot object detection at real-time speeds.

3.3.2 Image Segmentation

Image segmentation assigns a label to every pixel in an image. The main focus of early works was finding segments containing pixels with similar appearance or features, either with clustering-based methods [25, 26, 141] or graph-based algorithms [12, 37, 126, 158]. In deep learning frameworks, segmentation is typically performed using Fully Convolutional Network (FCN) architectures where encoders extract descriptive features and decoders generate per-pixel class predictions. In *semantic segmentation*, the network predicts an output vector per pixel, which is processed through a softmax layer to generate class probability distributions. The predicted classes are determined by taking the maxima of these probabilities. *Instance segmentation* differs from semantic segmentation by requiring differentiation between individual objects. In contrast to semantic segmentation, instance segmentation is usually solved by region-based methods closely related to object detectors. For example, Mask R-CNN [56] adds a segmentation head to a two-stage object detection network, which predicts pixel-wise binary masks for detected regions of interest. *Panoptic Segmentation* bridges these two concepts by proposing to predict both semantic and instance labels, and makes the distinction between *things* (countable objects) and *stuff* (uncountable regions such as sky, road or wall). Kirillov et al., after defining the task [71], proposed to solve it by adding a semantic segmentation branch to Mask R-CNN's backbone, in parallel to the existing instance segmentation head [70].

Recently, Kirillov et al. defined the new task of *promptable segmentation*, and simultaneously solved it with the Segment Anything (SAM) model [72]. SAM takes as input an image and a prompt, which can be a 2D bounding box, or one or more 2D coordinates. The prompts can be positive or negative, allowing for example users to manually click inside and outside objects of interest. As discussed in Sec. 3.2.8, SAM was trained on a very large dataset and consitutes the first foundation model for image segmentation. It is so capable that it can be straightforwardly put to use in different applications with practically no alteration. For example, it can be paired with an open-vocabulary object detector to form an open-vocabulary segmentation model [116]. This is the method we use in Paper IV to produce 2D object proposals in each view based on text prompts.

3.3.3 3D Object Detection

3D object detection involves localizing and classifying objects in three-dimensional scenes. In addition to the volumetric extent of the bounding box an optional rotation around the vertical axis can also be predicted. Some of the earliest deep learning methods targeted autonomous driving, where most interesting objects can be found in a planar Bird's Eye View (BEV) projection of the scene. By contrast, indoor scenes lack any simple global structure; objects vary in height and can appear anywhere in cluttered spaces, making such BEV-based assumptions invalid. Here we focus on indoor 3D object detection.

A major branch of 3D object detection uses point clouds or RGB-D images as input. Song et al. introduced a 3D Region Proposal Network to predict oriented bounding boxes in voxelized RGB-D data [130]. However, 3D CNNs on dense grids are computationally heavy and limited in resolution. VoteNet [111] leverages PointNet++ [110] to process point clouds directly without voxelization, taking advantage of the inherent sparsity of the data. It introduced a deep Hough voting mechanism for 3D object centers, achieving state-of-the-art results on indoor benchmarks using only geometric input. Meanwhile, voxel-based methods adopted sparse convolution networks that compute features only for occupied voxels, dramatically improving efficiency [121]. Recent approaches moved towards anchor-free and transformer-based architectures, employing set prediction techniques to directly output 3D boxes. For instance, Group-Free 3D [86] and 3DETR [100] predict boxes from a fixed number of learnable queries, analogously to DETR in 2D.

An alternative line of work performs 3D detection from multi-view RGB images, without requiring depth sensors. Inferring 3D structure from images is inherently ambiguous, but multi-view geometry can compensate for missing depth cues. ImVoxelNet is trained end-to-end on multiple views, projecting image features into a shared 3D voxel space and applying 3D convolutions to detect objects jointly across views [122]. ImGeoNet, building on ImVoxelNet, introduced an image-induced geometry-aware voxel representation that explicitly learns 3D shape cues from images [143]. ImGeoNet surpassed ImVoxelNet and even outperformed the point-cloud detector VoteNet in challenging scenarios, such as sparse or noisy point clouds or many small objects. Most recently, transformer-based models have emerged. For example, PARQ uses pixel-aligned 3D queries with cross-attention, achieving new state-of-the-art on indoor datasets using only RGB inputs [161].

All the above methods assume a fixed, closed set of object categories known a priori, which limits their ability to detect novel object types. The emerging field of open-vocabulary 3D detection (OV-3D) aims to lift this restriction by leveraging large vision-language models so that 3D detectors can recognize arbitrary classes. OV-3D remains challenging, due to the scarcity of large-scale 3D-text data and the difficulty of transferring 2D vision-language knowledge into 3D representations. Nevertheless, extending 3D detection to open-vocabulary is a necessary step toward general scene understanding. In Paper IV, we present a method for open-vocabulary 3D detection from sparse RGB views, bridging the gap between image-based geometry and open-vocabulary recognition.

3.4 Human Pose

The scientific study of human motion has roots that extend deep into history, reflecting humanity's fascination with understanding the mechanics of movement and the principles that govern bodily function. In Ancient Greece, Aristotle wrote the first documented treatise of biomechanical analysis [8], containing detailed observations of locomotion patterns of animals and humans, and proposing geometric principles of movement. During this period, artists also demonstrated remarkable understanding of human anatomy and strived to represent motion through the static medium of sculptures and frescoes.

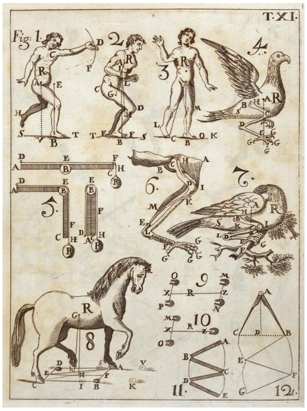


Figure 3.6: Giovanni Borelli's mechanical analysis of bones as levers [11].

In the 17th century, Giovanni Alfonso Borelli conducted formal experiments that established the mathematical foundations for understanding human motion. His work "De Motu Animalium" (On the Motion of Animals) [11] represents the first systematic attempt to apply mechanical principles the muscular system, analyzing muscle forces, measuring inhaled and exhaled air volumes and estimating with precision the energy exerted by each muscle. Borelli's insight that bones function as levers while muscles operate according to mathematical principles established fundamental concepts for modeling human motion that remain relevant in contemporary biomechanical analysis.

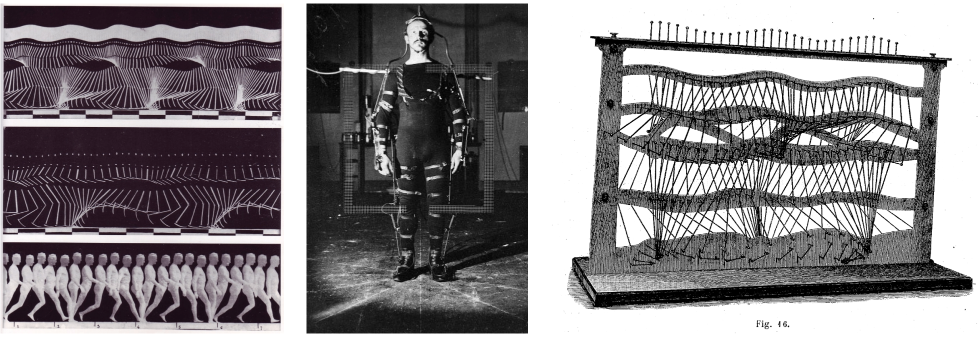


Figure 3.7: Chronophotography for motion analysis.. From left to right: Étienne-Jules Marey's late 19th-century kinograms, which decomposed continuous human motion into sequential visual frames, foreshadowing modern pose analysis; Braune and Fischer designed blinking markers enabling 3D reconstruction of movement from syncronized multi-view chronophotographs; the resulting 3D motion model [13].

The 19th century witnessed significant advances in the quantitative analysis of human locomotion, led by pioneers such as the Weber brothers, who conducted some of the first systematic studies of gait mechanics [155]. Their work established the importance of temporal analysis in understanding human movement, recognizing that the dynamics of motion contain as much information as static pose configurations.

In the late 19th century, physiologist Étienne-Jules Marey pioneered the use of graphical recording in the experimental sciences, designing instruments to record visually the evolution of physiological functions over time, such as the circulatory, respiratory and muscular systems [92]. Marey's development of *chronophotography*, inspired by Edweard Muybridge, represents a milestone in the visual analysis of motion. The technique of capturing multiple sequential images of moving subjects provided the first systematic method for decomposing continuous motion into discrete temporal samples that could be analyzed quantitatively. His experiments with subjects wearing black suits marked with white stripes along the limbs created visual representations, or *kinograms* that bear striking resemblance to the skeletal pose visualizations used in modern computer vision systems.

With quantitative studies came the first applications. Albert Londe, one of the first medical photographers, used chronophotography to study the movements of patients during epilectic seizures [87]. Christian Wilhelm Braune and Otto Fischer conducted experimental studies of human gait using blinking Geissler tube markers to record syncronized multi-view chronophotographs, enabling them to perform the first three-dimensional analysis of human motion [13].

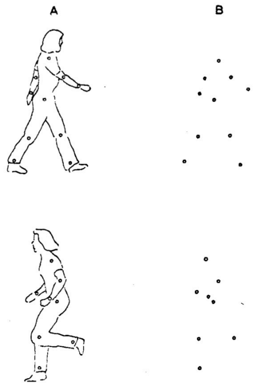


Figure 3.8: 2D keypoints (B) are sufficient to recognize complex actions [65].

Gunnar Johansson's pioneering studies of biological motion perception in the 1970's demonstrated that human observers could recognize complex actions from minimal visual information, using only point lights attached to major joints [65, 66]. This work showed that explicit modeling of body surface geometry and appearance details is not necessary for effective motion understanding, and that sparse skeletal representations were sufficient for conveying rich information about human movement. This finding supports the skeletal modeling approach used throughout this thesis, where human poses are represented as configurations of joint positions rather than detailed surface meshes or volumetric models.

The advent of digital computing enabled the development of marker-based motion capture systems that could provide precise 3D measurements of human movement in controlled environments. These systems established the gold standard for pose measurement accuracy and have been widely used in medical studies [14, 76, 104, 124], sports science [5, 134, 145] and movie production[125, 157]. However, the practical limitations of marker-based systems, including setup complexity, restricted capture volumes, and the need for specialized facilities [49, 101], motivated the development of markerless pose estimation methods that could operate from standard camera observations.

In this thesis, we focus on human pose estimation from images captured by multi-view cameras, which allows to capture subjects in their natural environment, such as workplaces, sport fields or public places, without encumbering their movements, without preparation, and without expensive rigs.

3.4.1 2D Human Pose Estimation

2D human pose estimation is the task of detecting the configuration of the human body in images or video frames. In practice, this entails localizing a set of predefined keypoints, usually corresponding to the main body joints such as the wrists, elbows, shoulders, knees, ankles, etc. Each person's keypoints can be connected in a skeletal structure to represent the pose.

Early approaches to 2D pose estimation relied on explicit graphical models of the human body, or pictorial structures. In pictorial structure models, introduced by Fischler and Elschlager [40], the human body is represented as a collection of rigid segments (limbs) connected by springs or flexible joints in a graphical model. The fundamental idea involves finding the optimal configuration of body parts that simultaneously minimizes appearance costs (how well each part matches the image evidence) and deformation costs (how much the configuration deviates from typical human anatomy).

The seminal work of Felzenszwalb and Huttenlocher [38] provided an efficient inference algorithm for tree-structured pictorial models using dynamic programming, making real-time pose estimation feasible for the first time. Their method formulated pose estimation as finding the optimal labeling of a graph where nodes represent body parts and edges encode spatial relationships. Yang and Ramanan extended this framework with the Flexible Mixture of Parts model [164], which learned different appearance templates for each body part under various viewpoints and deformations, significantly improving robustness to pose variation.

These classical approaches typically relied on hand-crafted features such as Histogram of Oriented Gradients (HOG) [27] or edge detection responses. While interpretable and often computationally efficient, these methods were fundamentally limited by their reliance on explicit modeling assumptions and their inability to capture the complex, non-linear relationships between appearance and pose that characterize real-world scenarios. Occlusions, particularly self-occlusions where parts of the person's own body occlude other parts, presented particularly challenging problems for these explicit models.

The emergence of deep learning, and particularly convolutional neural networks, marked a paradigm shift in human pose estimation. The availability of large-scale datasets such as MPII Human Pose [7] and MS COCO [83], combined with increased computational power, enabled the development of data-driven approaches that could learn complex appearance models directly from data.

Toshev and Szegedy introduced DeepPose [140], which represented the first successful application of deep neural networks to human pose estimation. DeepPose approached the problem as direct coordinate regression, learning a mapping from image pixels to joint coordinates through a deep convolutional neural network. The method employed a cascade of regressors, where initial pose estimates were iteratively refined by cropping regions around predicted joint locations and feeding them through subsequent networks. While computationally efficient, regression-based methods lack precision, particularly with complex poses and occlusions.

Detection-based methods, on the other hand, treat pose estimation as a spatial probability estimation problem. Rather than directly regressing coordinates,

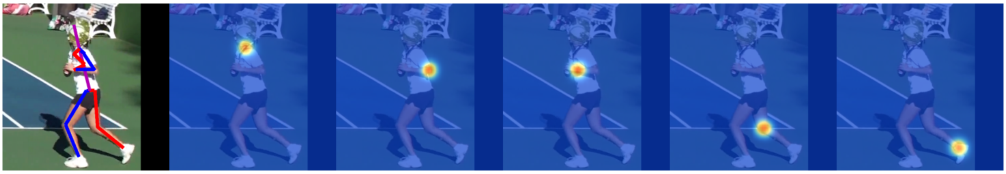


Figure 3.9: Heatmap-based 2D pose estimation. The model predicts one heatmap per joint (right). The final pose is given by the maximum activation for each heatmap (left). Image from [80].

these methods predict 2D heatmaps for each keypoint, representing the per-pixel likelihood for the joint positions, as shown in Fig. 3.9. The final pose is given by the maxima of the heatmaps. This formulation preserves spatial relationships and enables more robust optimization. Ground-truth heatmaps are typically generated by placing 2D Gaussians at each ground truth joint location. Tompson et al. introduced the heatmap approach with a multi-resolution network that predicted joint locations with a cascaded network combining coarse and fine heatmap regression [137, 138]. Wei et al. extended this concept with Convolutional Pose Machines (CPMs) [156], which sequentially refined joint predictions through multiple stages, with each stage having access to both image features and predictions from previous stages. The Stacked Hourglass Network combined an encoder-decoder structure with skip connections, enabling the network to capture both local detail and global context [80]. Multiple hourglass modules were stacked sequentially, with intermediate supervision applied at each stage, facilitating the learning of increasingly refined pose representations.

More recent architectural innovations have focused on improving the efficiency and accuracy of these foundational approaches. Xiao et al. demonstrated that simple baseline networks built on ResNet backbones with deconvolutional upsampling could achieve competitive performance, emphasizing the importance of strong feature representations [159]. HRNet (High-Resolution Network) maintained high-resolution feature maps throughout the network by parallel processing at multiple scales, thus addressing a limitation of previous networks that downsampled feature representations before upsampling them again [133, 151]. This architecture achieved state-of-the-art accuracy, particularly for precise joint localization.

3.4.2 Multi-Person Pose Estimation

Human pose estimation in images containing multiple persons presents additional challenges, such as inter-person occlusions, overlapping bodies, varying scales and, ultimately, the combinatorial complexity of assigning body joints to

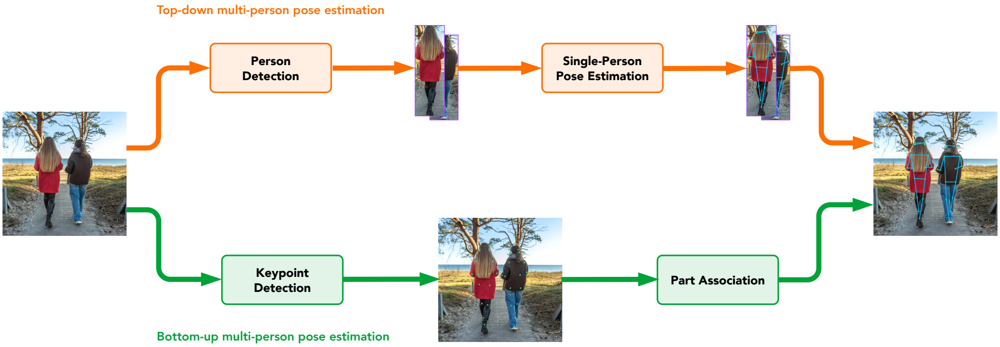

Top-down multi-person pose estimation
Person Detection
Single-Person Pose Estimation
Keypoint Detection
Part Association
Bottom-up multi-person pose estimation

Figure 3.10: Top-down pose estimation consists in detecting the persons in the image and performing single-person pose estimation. Bottom-up methods detect all the keypoints in the image and associating these keypoints to individual persons.

the correct persons. Multi-person pose estimation can in general be classified in two categories: *bottom-up* or *top-down* methods.

Bottom-up approaches detect all keypoints in the image simultaneously before solving the association problem to group keypoints into individual persons. Pishchulin et al. introduced DeepCut, the first deep learning-based bottom-up method, which formulated multi-person pose estimation as an integer linear program over detected body part candidates [108]. Cao et al. significantly advanced the paradigm with OpenPose, which introduced Part Affinity Fields (PAFs) to encode the spatial relationships between body parts [15, 16]. This efficient approach spearheaded real-time multi-person pose estimation. The computing speed of bottom-up approaches is generally unaffected by the number of people in the images. They can potentially recover people missed by person detectors and handle crowded scenarios gracefully. However, the association problem becomes increasingly challenging with large numbers of people, similar poses, or heavy occlusion, where the spatial cues encoded in PAFs may become ambiguous.

Top-down methods take a fundamentally different strategy and decompose the problem into two sequential stages, first performing person detection then estimating single-person poses within each detected bounding box. By isolating individual persons through bounding boxes, these methods can leverage the full context of the person region, making them accurate and robust. The modular approach allows each stage to be optimized independently, enabling to mix and match different flavours of state-of-the-art object detectors and single-person pose estimators depending on the application. One downside of top-down methods is that inference speed depends on the number of persons in the image.

However, in real-time applications with high enough frame rate, bounding boxes can be tracked with lightweight methods, alleviating the need to run person detection on each frame. Because of their flexibility and superior accuracy, we use top-down pose estimators in this thesis.

3.4.3 3D Human Pose Estimation

A natural extension of 2D pose estimation is to infer the spatial configuration of the body in three dimensions. This extension is motivated by applications requiring true spatial understanding of human movement, including biomechanical analysis, human-computer interaction, and augmented reality systems.

Monocular 3D Human Pose Estimation

Monocular 3D human pose estimation is one of the most challenging problems in computer vision, as it requires inferring three-dimensional structure from two-dimensional observations. The fundamental difficulty arises from the inherent ambiguity of perspective projection: infinitely many distinct 3D poses can produce an identical 2D projection on the image plane. This depth ambiguity, combined with other challenges such as self-occlusion and the variability of human appearance and motion, makes monocular 3D pose estimation an inherently ill-posed problem.

Early deep learning approaches addressed this task by directly regressing 3D pose coordinates from image pixels [106, 119, 135, 136]. However, such methods require large datasets of images with ground-truth 3D poses for training, typically obtained with specialized motion-capture systems, and tend to generalize poorly to images in the wild. Considerable effort has been devoted to overcoming the training data bottleneck. New datasets have been created by synthesizing realistic images of humans in known 3D poses, for example by rendering textured 3D human models [21], or by augmenting 2D pose datasets with plausible 3D annotations derived from mocap data [118]. Another strategy is to composite humans into natural images. Mehta et al. captured actors with a multi-camera markerless system in a green-screen studio and inserted them into in-the-wild backgrounds to produce training data with 3D ground truth [95, 96]. Alternatively, some methods eliminate explicit 3D labels by leveraging multiple camera views at training time. Rhodin et al. introduced a weakly-supervised approach that enforces cross-view consistency during training, i.e. the model is trained to predict the same 3D pose from different 2D views, thereby learning a monocular pose estimator without requiring a fully mocap-annotated dataset [117].

A common two-stage paradigm for monocular pose estimation is to first apply a 2D pose detector to the image and then *lift* the detected 2D joint coordinates to 3D. This approach was demonstrated effectively by Martinez et al. [94], whose *simple baseline* for monocular 3D pose estimation used an off-the-shelf 2D pose estimator [80] followed by a simple feed-forward network to predict 3D joint coordinates from 2D observations. Despite its simplicity, this two-step approach delivered competitive accuracy and underscored that much of the difficulty in monocular 3D pose estimation lies in resolving depth ambiguities rather than detecting 2D joint positions. Subsequent research has attempted to address the inherent ambiguity of the monocular problem through additional cues and constraints. Some methods exploit temporal information from video sequences, leveraging motion cues and enforcing temporal consistency to constrain the solution space [107]. Others incorporate strong priors about human anatomy and biomechanics, either through learned statistical models or explicit kinematic constraints. For example, Akhter and Black introduced pose-conditioned joint angle limits that rule out anatomically impossible configurations [6], while Wandt et al. proposed a kinematic chain space representation that naturally preserves constant limb lengths [150]. These model-based approaches regularize monocular predictions so that even in ambiguous cases, the reconstructed skeletons remain physically plausible. Another line of work uses adversarial training to encourage predicted poses to lie on the manifold of valid human poses. In these approaches, a discriminator is trained to differentiate between predicted poses and real human poses, and the pose estimator is penalized if its output is unrealistic [149].

Due to the inherent ambiguities of monocular views, monocular 3D pose estimation methods are typically evaluated under protocols that discount unknown global transformations. It is standard to measure error up to a rigid alignment (e.g. allowing rotation and scale adjustment via Procrustes analysis [47]) or up to a translation (by re-centering poses at a chosen root joint), acknowledging that absolute position cannot be reliably recovered from a single viewpoint. In practice, even state-of-the-art monocular methods struggle with extreme foreshortening or occlusion, as the missing depth information must be inferred from learned priors or temporal context rather than directly observed. This fundamental limitation motivates the use of multiple cameras, where geometric triangulation can resolve depth ambiguities that monocular methods can only guess.

Multi-view 3D pose estimation

With accurate camera calibration, the geometric relationships between views enable to recover absolute 3D coordinates without the scale and depth ambiguities inherent in single-view methods.

Triangulation-based methods represent the most straightforward approach to multi-view 3D pose estimation [58, 64, 90, 114, 160, 170]. These methods typically follow a two-stage pipeline, first estimating 2D poses in each view independently, then reconstructing 3D poses by triangulating corresponding keypoints with geometric algorithms such as the Direct Linear Transform (DLT) (c.f. Sec. 3.1.5) to find 3D points that minimize reprojection error across all views. However, in practice, 2D detections contain errors and may be inconsistent across views due to occlusion or lighting conditions, hence much attention has been given to improving the 2D keypoints. Several methods incorporate epipolar geometry into a 2D pose estimator's feature processing stage so that keypoint evidence from one view guides feature extraction in other views [58, 90, 170]. For example, the Epipolar Transformer uses attention along epipolar lines to refine feature maps across views [58], and AdaFuse adaptively weights the contributions of each view based on predicted per-view reliability, achieving robust results under occlusion [170]. Iskakov et al. implemented a weighted triangulation scheme accounting for detection confidence, which they use to supervise a 2D pose detector from multi-view data [64].

Volumetric methods construct a probabilistic 3D volume of the human body in voxel space. These approaches discretize the 3D capture volume and aggregate evidence from multiple views to create occupancy or probability volumes, from which human poses can be extracted. Iskakov et al. demonstrated that volumetric approaches could achieve high accuracy by projecting 2D CNN features into a discretized 3D volume and applying 3D convolutions to predict joint locations [64]. The volumetric representation naturally handles occlusion and provides spatial context, but computational costs scale cubically with volume resolution, limiting practical applications.

The main challenges for multi-view 3D pose estimation in real-world scenarios include handling occlusion across views, adapting to sparse camera configurations, and maintaining robustness when views have limited overlap. Many public datasets for multi-view pose estimation are characterized by dense camera coverage and controlled environments [63, 67], but practical deployments often involve wide-baseline setups with few cameras and significant occlusion.

Several works extend the multi-view paradigm to sequences, exploiting temporal

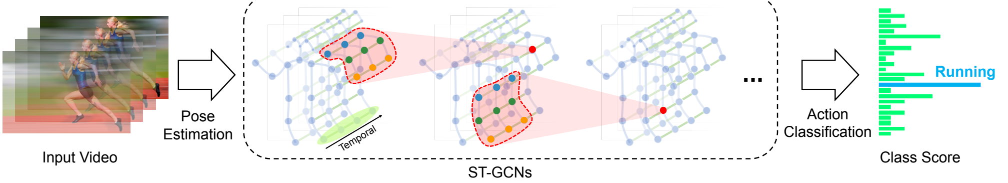


Figure 3.11: The ST-GCN pipeline. Image adapted from [162].

information to improve 3D pose reconstruction [28, 48, 129]. Temporal cues help smooth pose estimates, handle brief occlusions, and resolve ambiguities by enforcing motion coherence over time. However, these methods are generally not causal, i.e. they use both past and future frames for reconstruction, which is not suitable for real-time applications. In Paper III, we designed a Transformer encoder-decoder model to reconstruct 3D poses from sequences of multi-view 2D observations. In contrast to other methods leveraging temporal information, we focus on the causal case to enable real-time inference.

3.4.4 Skeleton-Based Action Recognition

The recognition of human activities from visual data represents a natural extension of human pose estimation. Activity recognition systems typically operate on sequences of pose observations, either in 2D image coordinates or 3D world coordinates, to classify the ongoing activity or predict future actions.

Traditional activity recognition approaches employed hand-crafted features based on trajectory analysis, optical flow, or space-time volumes. However, the advent of deep learning and the availability of large-scale datasets have enabled end-to-end learning approaches that can automatically discover discriminative spatio-temporal patterns.

Recognizing human activities from skeletal data offers several advantages over appearance-based methods. Skeletal representations are more compact than raw video, enabling efficient processing and storage. They are also more robust to variations in clothing, lighting, and background clutter that can confound appearance-based methods. Furthermore, as shown by Gunnar Johansson's experiments discussed above, skeletal encodes geometric relationships that, at least for humans, are sufficient for understanding human movement [65, 66].

The structural nature of human skeletons naturally suggests graph representations. Yan et al. revolutionized skeleton-based action recognition with Spatial-

Temporal Graph Convolutional Networks (ST-GCN), modeling skeletons as graphs where nodes represent joints and edges represent spatial (anatomical) or temporal (same joint across time) connections [162].

The ST-GCN model is composed of stacked spatio-temporal blocks. Each block performs a spatial convolution followed by a temporal convolution. Spatial convolution aggregates features from spatially neighbouring joints by graph convolution on the skeleton graph as described in c.f. Sec. 3.2.4. The temporal convolution submodule is a $1 \times T$ 2D convolution on the input, where T is the number of frames of the temporal receptive field. Thus ST-GCN captures spatial relationships between body parts and temporal evolution of the motion patterns. Note that the ST-GCN architecture can accomodate both 2D and 3D pose data, the only difference being the dimension of the input vectors.

Since its inception, several extensions to the original ST-GCN have been proposed. These include adaptive adjacency matrices that can learn task-specific graph structures beyond the fixed skeletal topology [127], attention mechanisms that dynamically weight the importance of different joints for specific activities [109], and multi-stream architectures that process different modalities of skeletal data (joint coordinates, bone vectors, motion information) in parallel before fusion [127, 128].

We use ST-GCN for activity recognition in Paper II, where we apply self-supervised learning techniques to learn robust action representations that can generalize across different domains with minimal labeled data.

Chapter 4

Research Summary and Conclusions

This thesis addresses key challenges encountered when designing and deploying multi-camera systems for 3D human and scene understanding beyond controlled laboratory settings. Paper I presents a practical method for extrinsic multi-camera calibration from 2D human poses enabling fast, equipment-free setup. In Paper II, we focus on learning generalizable representations for 3D skeleton-

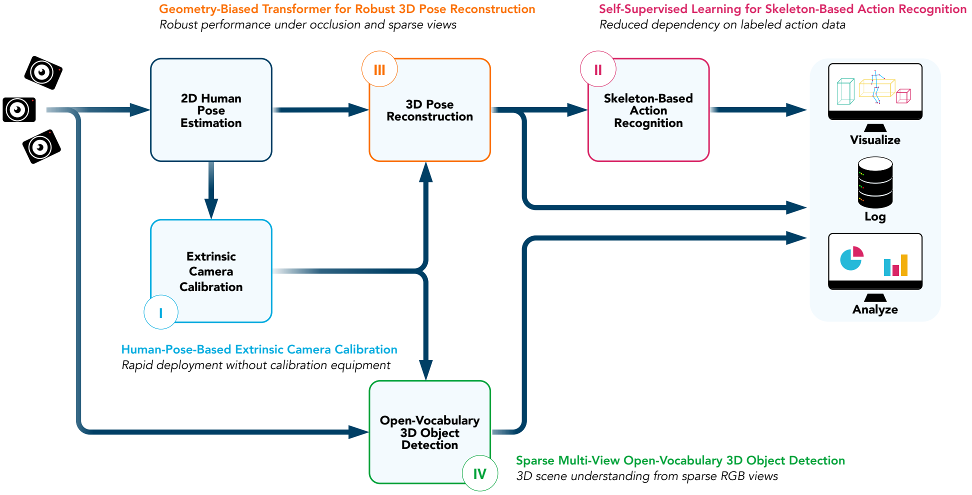

Geometry-Biased Transformer for Robust 3D Pose Reconstruction
Robust performance under occlusion and sparse views
Self-Supervised Learning for Skeleton-Based Action Recognition
Reduced dependency on labeled action data
2D Human Pose Estimation
III
3D Pose Reconstruction
II
Skeleton-Based Action Recognition
Visualize
Log
Analyze
Extrinsic Camera Calibration
I
Human-Pose-Based Extrinsic Camera Calibration
Rapid deployment without calibration equipment
Open-Vocabulary 3D Object Detection
IV
Sparse Multi-View Open-Vocabulary 3D Object Detection
3D scene understanding from sparse RGB views

Figure 4.1: Thesis contributions in the context of the MCS system.

based action recognition via self-supervised learning, reducing dependency on labeled training data. Paper III introduces a Transformer-based model for multi-view 3D human pose reconstruction that is robust to real-world challenges such as occlusions and sparse camera coverage. Finally, in Paper IV we propose a method for open-vocabulary 3D object detection from sparse multi-view RGB images, extending the system's capabilities to 3D scene understanding. Together, these works contribute to bringing multi-view computer vision from the laboratory to real-world deployment.

4.1 Human Pose-Based Camera Calibration (Paper I)

Traditional multi-camera calibration procedures rely on specialized equipment and skilled operators, hindering practical deployment of multi-camera systems. Establishing accurate point correspondences between views becomes particularly challenging in wide-baseline systems, where the appearance of the same 3D regions may vary dramatically across camera viewpoints due to occlusions, projective deformations, and viewpoint-dependent effects such as specular reflections. This makes the setup process of the system time-consuming and expensive, affecting every aspect of system deployment, from initial setup to ongoing maintenance. These were critical challenges for applications such as the traveling marketing experiences described in Sec. 2.2, which relied on rapid system deployment and reconfiguration across diverse venues.

Our key insight was given by the application context itself. Since our target scenarios involved detecting moving people, we speculated that the 2D joint detections could be used as correspondence points between camera views. Compared to classical feature descriptors, 2D joint detections are less dependent on local image appearance and more robust to viewpoint change, due to their semantic nature: like in X-ray imaging, the same anatomical joint can be seen ”through” the body, whereas the cameras may actually see two opposite sides of the body part with very different appearances. This makes human joints more reliable for correspondence matching across wide baselines compared to traditional feature descriptors.

Therefore, we developed a human-pose-based calibration method that estimates camera poses while reconstructing 3D human motion from 2D observations of a person moving naturally through the scene. This approach transforms the calibration process into a simple, intuitive task that requires only having someone walk around the capture space. The main challenge with this approach is that 2D joint detections are typically much less accurate than classical patch-based

2D image features.

After obtaining an initial estimate of the camera poses using standard Structure-from-Motion approaches with the potentially noisy keypoint detections, we propose several novel ideas to refine this estimate in the bundle adjustment step. First, we introduce a robust reprojection loss that takes into account the joint detection scores and the distances between the joints and the cameras. This allows the optimization to place greater emphasis on more reliable joint detections. Second, we implement a motion prior that encourages smooth joint trajectories while accounting for complex human motion, improving temporal consistency. Third, we enforce a constant limb length constraint that maintains anatomical consistency throughout the calibration sequence. Finally, we incorporate a learned 3D human pose likelihood model in kinematic chain space, effectively leveraging prior knowledge about human body configuration to prevent the optimization from converging to geometrically valid but biomechanically implausible solutions that might arise from noisy observations.

Our central hypothesis was that incorporating more domain-specific knowledge, i.e. understanding *how* 2d pose estimates are likely to be erroneous and *what* constitutes plausible human motion, could guide the calibration process towards more accurate and geometrically consistent solutions despite the noisy nature of the input data. Experimental validation on public datasets confirmed this hypothesis, demonstrating substantial improvements over previous human-pose-based calibration methods.

The calibration method presented in Paper I enables fast and accurate camera calibration while eliminating the need for specialized equipment. It has become a central part of the MCS system and has contributed to its success by enabling fast prototyping of new ideas and easy dissemination to remote teams without computer vision experience. It allows users to quickly set up and reconfigure a multi-camera system by simply having a person walk around the capture space, addressing the first research question (RQ1) of this thesis.

Performance depends on sufficient human motion to provide adequate geometric constraints for the optimization, and the approach relies fundamentally on the quality of 2D pose detection. When the pose detectors struggle due to poor lighting conditions, occlusions or challenging poses, the calibration accuracy can degrade significantly, requiring longer calibration sequences to be recorded. The current method still requires a user to deliberately walk around the capture area during setup or re-calibration. In some applications it would be practical to let the system calibrate itself using the poses from any and all the persons moving in the area; this would also enable the system to perform re-calibration automatically when certain conditions are detected. However, the people walking in the scene may follow paths that are not distributed ideally for calibration to succeed. This raises the question of how to automatically select which data is most useful for obtaining a good calibration result. Another interesting avenue for further research is to combine or initialize our method with learning-based relative pose regression methods that have recently made enormous progress in wide-baseline settings [32, 82, 152, 154, 169].

4.2 Learning Action Representations (Paper II)

In Paper II, we address the problem of developing effective action recognition systems without relying on large amounts of annotated data. As seen in Chapter 2, a key driving factor when developing the multi-camera system has been the ability to quickly prototype new ideas and address new needs. However, traditional supervised learning approaches to action recognition depend on large, carefully annotated datasets. Collecting and labeling action data proves both expensive and time-consuming, often requiring domain expertise to ensure consistent annotation quality. This becomes particularly problematic when exploring new application domains or when working with specialized activities that lack existing labeled datasets.

While labeled action data remains scarce and expensive to produce, unlabeled human motion data can be captured easily and continuously by multi-camera systems during normal operation. Inspired by the success of self-supervised learning in computer vision and natural language processing (see Sec. 3.2.7), we propose a self-supervised framework specifically designed for 3D pose sequence representation. Our method is based on Bootstrap Your Own Latent (BYOL) [50], a framework in which two neural networks, an online network and a target network, learn to predict consistent representations from two different augmented views of the same input data. The online network is trained to predict the representation produced by the target network, while the target network parameters are updated with an exponential moving average of the online network's weights, creating a sequence of online models of progressively increasing quality without any labeled data.

One contribution of Paper II is a data augmentation strategy for 3D pose sequence data that encourages the model to learn useful features while disregarding semantically-irrelevant variations. We developed a comprehensive set of skeleton-specific augmentations including temporal resampling to handle natural variations in action speed, spatial transformations that maintain anatomical structure while introducing viewpoint invariance, and filtering operations that simulate realistic noise perturbations.

While strong data augmentation contributes to making the model learn better features during pre-training, mild augmentations are preferred during fine-tuning for downstream tasks to avoid overwhelming the labeled training signal with excessive noise. This augmentation strategy mismatch can significantly impact transfer learning performance. To reduce this domain shift between the data seen when pre-training and during fine-tuning, Paper II proposes asymmetric augmentation pipelines that simultaneously expose the model to both aggressive and conservative augmentation distributions during pre-training. This asymmetric design encourages the model to learn representations that can map from heavily distorted sequences to representations that are consistent with those of minimally augmented sequences, which are preferred during fine-tuning. This effectively reduces the distribution gap between pre-training and transfer learning. We show that this leads to better performance in downstream tasks.

We also introduce multi-viewpoint sampling to leverage the fact that some datasets feature several recordings of the same pose sequences seen from different viewpoints. For example, the datasets used in the paper include synchronized sequences captured by several RGB-D cameras from different angles. Instead of treating these as separate, unrelated sequences, we sample pairs of different viewpoints of the same action as positive pairs during pre-training, which encourages the network to learn view-invariant representations and disregard artefacts of the depth camera or pose reconstruction system. Our experiments demonstrate that multi-viewpoint sampling significantly improves the quality of the learned representations. Note that this approach could be extended to data captured by RGB multi-camera systems: different 3D pose sequences can be reconstructed from the same motion sequence using 2D poses from different subsets of cameras.

Paper II directly addresses the third research question (RQ3). Effective representation learning of human motion data enables fine-tuning activity recognition models for downstream tasks with minimal labeled data. This makes the development and deployment of action recognition systems more accessible and adaptable to new application domains, significantly reducing the time and cost associated with data collection and annotation.

Effective pre-training requires access to sufficient quantities of unlabeled motion data, but more importantly to data that is varied enough to support generalization to downstream tasks. However, ensuring this diversity without prior knowledge of target domains remains challenging. Our approach demonstrates the most pronounced performance gains in low-data regimes, with benefits diminishing as labeled data becomes more abundant. On fully-labeled datasets, our method does not surpass fully-supervised baselines. This may however be a question of scale. The datasets used in our experiments, while standard in the field, are relatively small compared to the ”internet-scale” datasets that have driven breakthroughs in language models and vision foundation models. Recent work in computer vision has demonstrated that training large models on massive, diverse datasets can yield representations that substantially outperform smaller models when fine-tuned for downstream tasks. A compelling direction for future research would be to scale skeleton-based self-supervised learning to much larger datasets and model architectures. Multi-camera systems like MCS, deployed across diverse environments, could continuously collect unlabeled motion data at unprecedented scale. Training large foundation models on such datasets might yield representations that not only excel in few-shot scenarios but also surpass fully-supervised methods on standard benchmarks. Furthermore, such foundation models could enable a data engine approach similar to that pioneered by Segment Anything. A large, capable model could be used to (semi-)automatically annotate motion sequences, generating high-quality pseudo-labels for training smaller, deployment-ready models. This would address both the annotation bottleneck that currently limits the field and the computational constraints of real-time applications. The virtuous cycle of model improvement driving better automatic annotation, which in turn enables training even better models, could accelerate progress substantially while reducing the human effort required for new application domains.

4.3 Robust 3D Human Pose Estimation (Paper III)

In contrast to laboratory environments where camera placement, lighting conditions, and scene composition can be controlled, practical deployment of multi-camera system requires adapting to existing, often adverse conditions. Furniture, machinery or architectural features may create occlusions, and possible mounting locations are determined by the layout of the room, the availability of power outlets and the ease of running cables, resulting in areas with minimal camera view overlap. This makes it difficult to obtain accurate 3D pose reconstructions, as triangulation-based methods typically fail when observations are too noisy, or simply when keypoints are not visible in at least two views.

Whereas triangulation-based methods consider each 3D keypoint independently, the positions and motion patterns of the different parts of the body are deeply inter-related, reflecting anatomical connectivity, biomechanical principles, motor control patterns, and temporal continuity constraints. In Paper III, we approach multi-view 3D human pose reconstruction as a data-driven regression problem, and designed a novel encoder-decoder Transformer model to uncover these underlying relationships between 3D joints from multi-view 2D observations.

The first component of our model is a Transformer encoder that takes as input 2D joint detections encoded as 3D rays passing through the camera centers and the detected joints, using a Plücker representation that decouples ray coordinates from camera positions. The encoder treats all these joints detected in different views and at different times as individual tokens and processes them globally, allowing information to flow across views, joints and time, even when some joints are not consistently visible. Whereas standard Transformers treat all tokens equally, we introduce a biased attention mechanism to incorporate domain-specific knowledge about the reliability of the observations: a confidence bias weigths attention based on the detection confidence scores from the 2D pose estimator, and a geometry bias promotes attention between rays that are close in 3D space. This effectively guides the attention mechanism towards observations that are both reliable and geometrically consistent. The sequence of refined tokens produced by the encoder form a global representation of the pose sequence.

The second component of the model is a 3D pose sequence decoder that queries the global encoded representation using predefined joint queries encoding both semantic information about specific body parts and temporal information about target time frames. These queries allow to extract specific 3D pose information from the global encoded representation, enabling flexible output generation where the number of input and output frames can differ.

To promote generalization to unseen scenes and improve robustness to missing joint observations, we implement several training strategies. Scene centering transforms all observations to a subject-centric coordinate system, allowing the model to handle scenes of varying dimensions. We generate synthetic views during training to increase the diversity of camera poses, compensating for the limited viewpoint variations of existing datasets and improving generalization to novel camera configurations. We also leveraged token dropout, which randomly removes input tokens during training, emulating real-world occlusion scenarios where certain body parts may be temporarily invisible due to furniture, other people, or self-occlusion, to make the model resilient to missing joints and incomplete temporal sequences.

The contributions of Paper III address our second research question (RQ2) by providing a robust solution for 3D human pose reconstruction in situations with occlusions and limited camera views. This enables robust performance in diverse real-world deployment scenarios where the environment cannot be controlled, advancing the practical applicability of multi-camera systems such as MCS.

The performance of our method can degrade for poses that differ substantially from the training distribution, which highlights the importance of diverse and representative training datasets. Unfortunately, datasets combining multi-view videos and annotated 3D pose data are few, relatively small and do not offer much diversity of movements. Systems such as MCS, which can be easily installed in various environments, enable continuous gathering of multi-view 2D pose data and automatically triangulated 3D poses. However, curating the 3D poses to ensure that they have the quality required to train a proficient pose reconstruction model still represents a considerable amount of work. It would be interesting to explore weakly-supervised learning approaches for training reconstruction models on unlabeled multi-view 2D data, which would enable to train on substantially larger datasets without 3D annotations. Another strategy would be to pre-train the networks on large-scale 3D motion capture datasets such as [91], which do not provide 2D data. Finally, recent work has shown that leveraging physics priors, e.g. via simulation, enables the reconstruction of more plausible poses [42, 79, 142]. An interesting future research direction could be exploring physics simulation during training, or for automatic ground-truth generation from unlabeled multi-view 2D data.

4.4 3D Scene Understanding from Sparse 2D Views (Paper IV)

Papers II and III focus on human pose estimation and analysis, without considering the environment in which the persons are evolving. In Paper IV, we explore whether the same system of sparse, fixed 2D cameras can be used to achieve a 3D understanding of the scene, i.e. where are the walls, the furniture, what objects are in the room, and where they are located. To that end, we tackle the problem of open-vocabulary 3D object detection from sparse multiview 2D images.

Our method is training-free and relies entirely on pre-trained, off-the-shelf 2D networks, making it immediately applicable without requiring domain-specific fine-tuning. The first stage generates initial 3D object proposals from individual RGB images by applying a state-of-the-art open-vocabulary 2D object detector [98] to identify objects specified by text queries, then using an image segmentation model [72] to obtain accurate 2D masks for each detected object. We lift the 2D masks to 3D using monocular depth estimation using MoGe [153], an affine-invariant monocular depth estimator that predicts relative depth maps. We backproject each pixel within an object mask to 3D space using the estimated depth value and known camera parameters. This stage produces a set of 3D point clouds, one for each detected object in each view.

However, these initial proposals suffer from the scale ambiguity inherent in monocular depth estimation and may have inconsistent depth scales across different views or even within individual images. The core contribution of our approach lies in the multi-view refinement stage, which optimizes the initial 3D proposals for consistency across different camera views. Before optimizing individual proposals, we estimate a global scale factor that provides a reasonable initialization for all proposals in a given view. This addresses the fact that monocular depth estimators often produce depth maps with globally consistent relative scales, even if the absolute scale is unknown. We then optimize individual scale and shift parameters for each object proposal independently. This per-proposal refinement accounts for local inconsistencies in the monocular depth maps that cannot be corrected by a global scale factor alone. Objects at different distances from the camera, or in different parts of the image, may require different scale adjustments to achieve multi-view consistency. We define a multi-view consistency loss that measures how well a 3D proposal aligns with the corresponding image content when projected into other camera views. The loss combines two complementary terms: the first term encourages photometric alignment across views, while the second term captures semantic consistency and is more robust to photometric variations due to lighting changes or viewpoint-dependent effects. The combination of these terms provides a robust measure of multi-view consistency that can handle photometric variations across views.

The final stage of our pipeline combines the optimized 3D proposals from different views into coherent 3D object detections by greedily merging the axis-aligned 3D bounding boxes computed for each optimized proposal, based on the intersection-over-union (IoU) of the boxes. For each merged cluster, we compute a final bounding box that encompasses the union of all point clouds belonging to proposals in that cluster.

A key strength of our method is that it relies entirely on pre-trained, off-the-shelf 2D networks without any 3D-specific training, resulting in better generalization to new object categories and environments. The training-free nature also enables immediate deployment without the data collection and training overhead associated with supervised 3D approaches.

Paper IV addresses the fourth research question (RQ4) by providing comprehensive scene understanding from few 2D views, thus making it possible to leverage the same system that was designed for real-time human pose reconstruction and analysis.

This work represents an initial step toward comprehensive 3D scene understanding from sparse views. A natural next step would be to explore human-object interaction analysis by combining the 3D human poses from our multi-view system with detected object locations and affordances. This integration could enable activity recognition that leverages both motion patterns and environmental context, for example distinguishing between ”reaching for a cup” and ”reaching for a book” based on the spatial relationship between the person and nearby objects. Knowledge of the 3D scene could also be leveraged to enhance 3D human pose reconstruction, by preventing unrealistic collisions and intersections with the environment. Another promising direction is to use the semantic and geometric data provided by our method to reason about object relationships, spatial layouts, and functional affordances, enabling natural language queries about complex spatial relationships, e.g. ”Where is the safest place to put this fragile item?”. Beyond static object detection, the framework could be extended to understand dynamic scene properties, including extrapolating the state of the scene in invisible areas, modeling temporal scene evolution, and predicting probable future states (”will that precariously placed item fall?”). Such capabilities would enable proactive assistance systems that can anticipate events before they occur.

References

[1] https://www.hawkeyeinnovations.com. 2

[2] https://www.optitrack.com. 1

[3] https://www.qualisys.com. 1

[4] https://www.vicon.com. 1

[5] B. Adlou, C. Wilburn, and W. Weimar. Motion capture technologies for athletic performance enhancement and injury risk assessment: A review for multi-sport organizations. *Sensors*, 25(14), 2025. ISSN 1424-8220. doi: 10.3390/s25144384. URL https://www.mdpi.com/1424-8220/25/14/4384. 1, 40

[6] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1446–1455, 2015. doi: 10.1109/CVPR.2015.7298751. 45

[7] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014. doi: 10.1109/CVPR.2014.471. 41

[8] Aristotle. De motu animalium; de incessu animalium. In W. D. Ross, J. A. Smith, et al., editors, *The works of Aristotle*, chapter V. Clarendon Press, Oxford, 1912. [ca. 350 BCE; trad. A. S. L. Farquharson]. 38

[9] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. In *European Conference on Computer Vision (ECCV)*, pages 404–417. Springer, 2006. 16

[10] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 33

[11] G. A. Borelli. *De motu animalium.* Ex typographia Angeli Bernabo, Romae, 1680. 38

[12] Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 1, pages 105–112 vol.1, 2001. doi: 10.1109/ICCV.2001.937505. 36

[13] W. Braune and O. Fischer. *Der Gang des Menschen.* Number I in Abhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften [zu Leipzig]. B.S. Hirzel, 1895. 39

[14] C. Buckley, L. Alcock, R. McArdle, R. Z. U. Rehman, S. Del Din, C. Mazzà, A. J. Yarnall, and L. Rochester. The role of movement analysis in diagnosing and monitoring neurodegenerative conditions: Insights from gait and postural control. *Brain Sciences*, 9(2), 2019. ISSN 2076-3425. doi: 10.3390/brainsci9020034. URL https://www.mdpi.com/2076-3425/9/2/34. 1, 40

[15] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310, 2017. doi: 10.1109/CVPR.2017.143. 1, 43

[16] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1): 172–186, 2021. doi: 10.1109/TPAMI.2019.2929257. 43

[17] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-End object detection with transformers. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing. 28, 35, 36

[18] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in Self-Supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 32

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. Feb. 2020. 32

[20] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton. Big Self-Supervised models are strong Semi-Supervised learners. June 2020. 32

[21] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing Training Images for Boosting Human 3D Pose Estimation . In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 479–488, Los Alamitos, CA, USA, Oct. 2016. IEEE Computer Society. doi: 10.1109/3DV.2016.58. URL https://doi.ieeecomputersociety.org/10.1109/3DV.2016.58. 44

[22] X. Chen and K. He. Exploring simple siamese representation learning. Nov. 2020. 32

[23] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. Mar. 2020. 32

[24] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan. Yolo-world: Real-time open-vocabulary object detection. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16901–16911, 2024. 36

[25] G. Coleman and H. Andrews. Image segmentation by clustering. *Proceedings of the IEEE*, 67(5):773–785, 1979. doi: 10.1109/PROC.1979.11327. 36

[26] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002. doi: 10.1109/34.1000236. 36

[27] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005. doi: 10.1109/CVPR.2005.177. 34, 41

[28] V. Davoodnia, S. Ghorbani, M.-A. Carbonneau, A. Messier, and A. Etemad. Upose3d: Uncertainty-aware 3d human pose estimation with cross-view and temporal cues. In *European Conference on Computer Vision*, 2024. 47

[29] M. Demazure. Sur deux problèmes de reconstruction. Technical Report RR-0882, INRIA, July 1988. URL https://inria.hal.science/inria-00075672. 19

[30] D. DeTone, T. Malisiewicz, and A. Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 224–236, 2018. 16

[31] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015. 32

[32] S. Dong, S. Wang, S. Liu, L. Cai, Q. Fan, J. Kannala, and Y. Yang. Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 16739–16752, June 2025. 52

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 28

[34] J. Edstedt. Less Biased Noise Scale Estimation for Threshold-Robust RANSAC. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2025. 20

[35] J. Edstedt, G. Bökman, M. Wadenbäck, and M. Felsberg. Dedode: Detect, don't describe — describe, don't detect for local feature matching. In *2024 International Conference on 3D Vision (3DV)*, pages 148–157, 2024. doi: 10.1109/3DV62453.2024.00035. 16

[36] S. Edunov, M. Ott, M. Auli, and D. Grangier. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1045. URL https://aclanthology.org/D18-1045. 28

[37] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004. doi: 10.1023/B:VISI.0000022288.19776.77. URL https://doi.org/10.1023/B:VISI.0000022288.19776.77. 36

[38] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005. doi: 10.1023/B:VISI.0000042934.15159.49. URL https://doi.org/10.1023/B:VISI.0000042934.15159.49. 41

[39] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 20

[40] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, C-22:67–92, 1973. 40

[41] R. Fletcher. *A modified Marquardt subroutine for non-linear least squares*. AERE report / R.: AERE report. Atomic Energy Research Establishment, Harwell, 1971. 23

[42] E. Gartner, M. Andriluka, E. Coumans, and C. Sminchisescu. Differentiable dynamics for articulated 3d human motion reconstruction. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13180–13190, 2022. 56

[43] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, ICLR*, 2018. URL https://openreview.net/forum?id=S1v4N2l0-. 32

[44] R. Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. doi: 10.1109/ICCV.2015.169. 34

[45] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. doi: 10.1109/CVPR.2014.81.

[46] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1): 142–158, 2016. doi: 10.1109/TPAMI.2015.2437384. 34

[47] C. Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(2):285–321, 12 2018. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1991.tb01825.x. URL https://doi.org/10.1111/j.2517-6161.1991.tb01825.x. 45

[48] B. Gordon, S. Raab, G. Azov, R. Giryes, and D. Cohen-Or. Flex: Extrinsic parameters-free multi-view 3d human motion reconstruction. In *European Conference on Computer Vision*, 2021. 47

[49] G. E. Gorton, D. A. Hebert, and M. E. Gannotti. Assessment of the kinematic variability among 12 motion analysis laboratories. *Gait & Posture*, 29(3):398–402, 2009. ISSN 0966-6362. doi: https://doi.org/10.1016/j.gaitpost.2008.10.060. 1, 40

[50] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent - a new approach to Self-Supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. 27, 32, 52

[51] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=lL3lnMbR4WU. 35

[52] R. Hartley. Projective reconstruction and invariants from multiple images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(10):1036–1041, 1994. doi: 10.1109/34.329005. 19

[53] R. Hartley. In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):580–593, 1997. doi: 10.1109/34.601246. 19

[54] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004. second edition. 20, 21

[55] R. I. Hartley and P. Sturm. Triangulation. *Computer Vision and Image Understanding*, 68(2):146–157, 1997. ISSN 1077-3142. doi: https://doi.org/10.1006/cviu.1997.0547. URL https://www.sciencedirect.com/science/article/pii/S1077314297905476. 21

[56] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. doi: 10.1109/ICCV.2017.322. 34, 36

[57] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020. 32

[58] Y. He, R. Yan, K. Fragkiadaki, and S.-I. Yu. Epipolar transformers. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 7779–7788, 2020. 46

[59] S. B. Heinrich. Efficient and robust model fitting with unknown noise scale. *Image Vision Comput.*, 31(10):735–747, Oct. 2013. ISSN 0262-8856. doi: 10.1016/j.imavis.2013.07.003. URL https://doi.org/10.1016/j.imavis.2013.07.003. 20

[60] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(89)90020-8. URL https://www.sciencedirect.com/science/article/pii/0893608089900208. 24

[61] N. Högberg, D. Berthet, M. Alam, P. P. Nielsen, L.-M. Tamminen, N. Fall, and A. Kroese. Exploring pose estimation as a tool for the assessment of brush use patterns in dairy cows. *Applied Animal Behaviour Science*, 292:106746, 2025. ISSN 0168-1591. doi: https://doi.org/10.1016/j.applanim.2025.106746. URL https://www.sciencedirect.com/science/article/pii/S0168159125002448. 9

[62] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/ioffe15.html. 27

[63] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, July 2014. 2, 46

[64] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov. Learnable triangulation of human pose. In *International Conference on Computer Vision (ICCV)*, 2019. 22, 46

[65] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, 1973. doi: 10.3758/BF03212378. URL https://doi.org/10.3758/BF03212378. 40, 47

[66] G. Johansson. Spatio-temporal differentiation and integration in visual motion perception. *Psychological Research*, 38(4):379–393, 1976. doi: 10.1007/BF00309043. URL https://doi.org/10.1007/BF00309043. 40, 47

[67] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. S. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2, 46

[68] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 31

[69] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017. 28

[70] A. Kirillov, R. Girshick, K. He, and P. Dollár. Panoptic feature pyramid networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6392–6401, 2019. doi: 10.1109/CVPR.2019.00656. 36

[71] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. Panoptic segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9396–9405, 2019. doi: 10.1109/CVPR.2019.00963. 36

[72] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, October 2023. 34, 36, 57

[73] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf. 24

[74] A. Kroese, M. Alam, E. Hernlund, D. Berthet, L.-M. Tamminen, N. Fall, and N. Högberg. 3-dimensional pose estimation to detect posture transition in freestall-housed dairy cows. *Journal of Dairy Science*, 107 (9):6878–6887, 2025/07/28 2024. doi: 10.3168/jds.2023-24427. URL https://doi.org/10.3168/jds.2023-24427. 9

[75] A. Kroese, N. Högberg, E. Diaz Vicuna, D. Berthet, N. Fall, M. Alam, and L.-M. Tamminen. Evaluating the automated measurement of abnormal rising and lying down behaviours in dairy cows using 3d pose estimation. *Smart Agricultural Technology*, 12:101205, 2025. ISSN 2772-3755. doi: https://doi.org/10.1016/j.atech.2025.101205. URL https://www.sciencedirect.com/science/article/pii/S2772375525004368. 9

[76] W. W. T. Lam, Y. M. Tang, and K. N. K. Fong. A systematic review of the applications of markerless motion capture (mmc) technology for clinical measurement in rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, 20(1):57, 2023. doi: 10.1186/s12984-023-01186-9. URL https://doi.org/10.1186/s12984-023-01186-9. 1, 40

[77] G. Larsson, M. Maire, and G. Shakhnarovich. Learning representations for automatic colorization. In *Computer Vision – ECCV 2016*, pages 577–593. Springer International Publishing, 2016. 32

[78] V. Larsson. *Computational Methods for Computer Vision: Minimal Solvers and Convex Relaxations*. Doctoral thesis (monograph), Lund University, 2018. 21

[79] C. Le, V. Johansson, M. Kok, and B. Wandt. Optimal-state dynamics estimation for physics-based human motion capture from videos. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 43609–43631. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4cf53c9fa86318d37bbeadac59c1a34d-Paper-Conference.pdf. 56

[80] B. Leibe, J. Matas, N. Sebe, and M. Welling, editors. *Stacked Hourglass Networks for Human Pose Estimation*, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46484-8. 1, 42, 45

[81] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o(n) solution to the pnp problem. *Int. J. Comput. Vision*, 81(2):155–166, Feb. 2009. ISSN 0920-5691. doi: 10.1007/s11263-008-0152-6. URL https://doi.org/10.1007/s11263-008-0152-6. 23

[82] V. Leroy, Y. Cabon, and J. Revaud. Grounding image matching in 3d with mast3r, 2024. 52

[83] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context.

In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. 41

[84] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. yue Li, J. Yang, H. Su, J.-J. Zhu, and L. Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2023. URL https://api.semanticscholar.org/CorpusID:257427307. 36

[85] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46448-0. 35

[86] Z. Liu, Z. Zhang, Y. Cao, H. Hu, and X. Tong. Group-free 3d object detection via transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2929–2938, 2021. doi: 10.1109/ICCV48922.2021.00294. 37

[87] A. Londe. *La photographie médicale: application aux sciences médicales et physiologiques*. Gauthier-Villars, 1893. 39

[88] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293(5828):133–135, 1981. 19

[89] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 16

[90] H. Ma, L. Chen, D. Kong, Z. Wang, X. Liu, H. Tang, X. Yan, Y. Xie, S.-Y. Lin, and X. Xie. Transfusion: Cross-view fusion with transformer for 3d human pose estimation. In *British Machine Vision Conference*, 2021. 46

[91] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. Black. Amass: Archive of motion capture as surface shapes. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5441–5450, 2019. doi: 10.1109/ICCV.2019.00554. 56

[92] É. J. Marey. *La méthode graphique dans les sciences expérimentales et particulièrement en physiologie et en médecine*. G. Masson, 1878. 39

[93] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963. 23

[94] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2659–2668, 2017. doi: 10.1109/ICCV.2017.288. 45

[95] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 International Conference on 3D Vision (3DV)*, pages 506–516, 2017. doi: 10.1109/3DV.2017.00064. 44

[96] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3D Vision (3DV), 2018 Sixth International Conference on*. IEEE, sep 2018. URL http://gvv.mpi-inf.mpg.de/projects/SingleShotMultiPerson. 44

[97] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer, 2022. 35

[98] M. Minderer, A. A. Gritsenko, and N. Houlsby. Scaling open-vocabulary object detection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=mQPNcBWjGc. 35, 56

[99] I. Misra and L. van der Maaten. Self-supervised learning of pretext-invariant representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020. 32

[100] I. Misra, R. Girdhar, and A. Joulin. An end-to-end transformer model for 3d object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2886–2897, 2021. doi: 10.1109/ICCV48922.2021.00290. 37

[101] L. Mündermann, S. Corazza, and T. P. Andriacchi. The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications. *Journal of NeuroEngineering and Rehabilitation*, 3(1):6, 2006. doi: 10.1186/1743-0003-3-6. URL https://doi.org/10.1186/1743-0003-3-6. 1, 40

[102] D. Nister. An efficient solution to the five-point relative pose problem. In *2003 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–195. IEEE, 2003. 19

[103] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision - ECCV 2016*, volume 9910 of *Lecture Notes in Computer Science*, pages 69–84. Springer, 2016. ISBN 978-3-319-46465-7. 32

[104] M. Osborne, N. M. Mueske, S. A. Rethlefsen, R. M. Kay, and T. A. Wren. Pre-operative hamstring length and velocity do not explain the reduced effectiveness of repeat hamstring lengthening in children with cerebral palsy and crouch gait. *Gait & Posture*, 68:323–328, 2019. ISSN 0966-6362. doi: https://doi.org/10.1016/j.gaitpost.2018.11.033. URL https://www.sciencedirect.com/science/article/pii/S0966636218307434. 1, 40

[105] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 32

[106] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1263–1272, 2017. doi: 10.1109/CVPR.2017.139. 44

[107] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7745–7754, 2019. doi: 10.1109/CVPR.2019.00794. 45

[108] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation . In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4929–4937, Los Alamitos, CA, USA, June 2016. IEEE Computer Society. doi: 10.1109/CVPR.2016.533. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.533. 43

[109] C. Plizzari, M. Cannici, and M. Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *Comput. Vis. Image Underst.*, 208-209:103219, 2020. 48

[110] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the*

31st International Conference on Neural Information Processing Systems, NIPS'17, page 5105–5114, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964. 37

[111] C. R. Qi, O. Litany, K. He, and L. J. Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 37

[112] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html. 33, 35

[113] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 35

[114] E. Remelli, S. Han, S. Honari, P. Fua, and R. Wang. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6039–6048, jun 2020. doi: 10.1109/CVPR42600.2020.00608. URL https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00608. 46

[115] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 34, 35

[116] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 36

[117] H. Rhodin, F. Meyer, J. Spörri, E. Müller, V. Constantin, P. Fua, I. Katircioglu, and M. Salzmann. Learning monocular 3d human pose estimation from multi-view images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8437–8446, 2018. doi: 10.1109/CVPR.2018.00880. 45

[118] G. Rogez and C. Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3116–3124, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819. 44

[119] G. Rogez, P. Weinzaepfel, and C. Schmid. Lcr-net: Localization-classification-regression for human pose. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1216–1224, 2017. doi: 10.1109/CVPR.2017.134. 44

[120] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski. ORB: An efficient alternative to SIFT or SURF. In *2011 International Conference on Computer Vision (ICCV)*, pages 2564–2571. IEEE, 2011. 16

[121] D. Rukhovich, A. Vorontsova, and A. Konushin. Fcaf3d: Fully convolutional anchor-free 3d object detection. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision – ECCV 2022*, pages 477–493, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20080-9. 37

[122] D. Rukhovich, A. Vorontsova, and A. Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1265–1274, 2022. doi: 10.1109/WACV51458.2022.00133. 37

[123] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. doi: 10.1038/323533a0. URL `https://doi.org/10.1038/323533a0`. 30, 31

[124] F. Salami, J. Brosa, S. Van Drongelen, M. C. M. Klotz, T. Dreher, S. I. Wolf, and M. Thielen. Long-term muscle changes after hamstring lengthening in children with bilateral cerebral palsy. *Developmental Medicine & Child Neurology*, 61(7):791–797, 2019. doi: https://doi.org/10.1111/dmcn.14097. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/dmcn.14097`. 1, 40

[125] S. Sharma, S. Verma, M. Kumar, and L. Sharma. Use of motion capture in 3d animation: Motion capture systems, challenges, and recent trends. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 289–294, 2019. 1, 40

[126] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. doi: 10.1109/34.868688. 36

[127] L. Shi, Y. Zhang, J. Cheng, and H. Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12018–12027, 2018. 48

[128] L. Shi, Y. Zhang, J. Cheng, and H. Lu. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29:9532–9545, 2019. 48

[129] H. Shuai, L. Wu, and Q. Liu. Adaptive multi-view and temporal fusing transformer for 3d human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:4122–4135, 2021. 47

[130] S. Song and J. Xiao. Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images . In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 808–816, Los Alamitos, CA, USA, June 2016. IEEE Computer Society. doi: 10.1109/CVPR.2016.94. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.94. 37

[131] P. Stefanovic. Relative orientation–a new approach. *ITC Journal*, 3(417-448):2, 1973. 19

[132] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou. Loftr: Detector-free local feature matching with transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8918–8927, 2021. doi: 10.1109/CVPR46437.2021.00881. 16

[133] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 1, 42

[134] X. Suo, W. Tang, and Z. Li. Motion capture technology in sports scenarios: A survey. *Sensors*, 24(9), 2024. ISSN 1424-8220. doi: 10.3390/s24092947. URL https://www.mdpi.com/1424-8220/24/9/2947. 1, 40

[135] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua. Structured prediction of 3d human pose with deep neural networks. In *Procedings of the British Machine Vision Conference 2016*, pages 130–1. British Machine Vision Association, 2016. 44

[136] B. Tekin, P. Marquez Neila, M. Salzmann, and P. Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3961–3970, 2017. doi: 10.1109/ICCV.2017.425. URL https://infoscience.epfl.ch/handle/20.500.14299/139835. 44

[137] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, pages 1799–1807, Cambridge, MA, USA, 2014. MIT Press. 1, 42

[138] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 648–656, 2015. doi: 10.1109/CVPR.2015.7298664. 42

[139] M. Topley and J. G. Richards. A comparison of currently available optoelectronic motion capture systems. *Journal of Biomechanics*, 106:109820, 2020. ISSN 0021-9290. doi: https://doi.org/10.1016/j.jbiomech.2020.109820. URL https://www.sciencedirect.com/science/article/pii/S0021929020302438. 1

[140] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, 2013. 1, 41

[141] J. Tou and R. Gonzalez. *Pattern Recognition Principles*. Applied mathematics and computation. Addison-Wesley Publishing Company, 1974. ISBN 9780201075878. 36

[142] S. Tripathi, L. Müller, C.-H. P. Huang, O. Taheri, M. J. Black, and D. Tzionas. 3d human pose estimation via intuitive physics. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4713–4725, 2023. doi: 10.1109/CVPR52729.2023.00457. 56

[143] T. Tu, S.-P. Chuang, Y.-L. Liu, C. Sun, K. Zhang, D. Roy, C.-H. Kuo, and M. Sun. Imgeonet: Image-induced geometry-aware voxel representation for multi-view 3d object detection. In *Proceedings of the IEEE international conference on computer vision*, 2023. 37

[144] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. July 2018. 32

[145] E. van der Kruk and M. M. Reijne. Accuracy of human motion capture systems for sport applications; state-of-the-art review. *European Journal of Sport Science*, 18(6):806–819, 2018. doi: https://doi.org/10.1080/17461391.2018.1463397. URL https://onlinelibrary.wiley.com/doi/abs/10.1080/17461391.2018.1463397. 1, 40

[146] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. U. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 28

[147] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 1096–1103, New York, NY, USA, July 2008. Association for Computing Machinery. 32

[148] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001. doi: 10.1109/CVPR.2001.990517. 34

[149] B. Wandt and B. Rosenhahn. RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation . In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7774–7783, Los Alamitos, CA, USA, June 2019. IEEE Computer Society. doi: 10.1109/CVPR.2019.00797. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00797. 45

[150] B. Wandt, H. Ackermann, and B. Rosenhahn. A kinematic chain space for monocular motion capture. In L. Leal-Taixé and S. Roth, editors, *Computer Vision – ECCV 2018 Workshops*, pages 31–47, Cham, 2019. Springer International Publishing. ISBN 978-3-030-11018-5. 45

[151] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019. 42

[152] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 52

[153] R. Wang, S. Xu, C. Dai, J. Xiang, Y. Deng, X. Tong, and J. Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5261–5271, June 2025. 34, 57

[154] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 52

[155] W. Weber and E. F. Weber. *Mechanik der menschlichen Gehwerkzeuge: eine anatomisch-physiologische Untersuchung*, volume 1. Dietrich, 1836. 39

[156] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016. doi: 10.1109/CVPR.2016.511. 1, 42

[157] M. C. Wibowo, S. Nugroho, and A. Wibowo. The use of motion capture technology in 3d animation. *International Journal of Computing and Digital Systems*, 2024. 1, 40

[158] Z. Wu and R. Leahy. An optimal graph theoretic approach to data clustering: theory and its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1101–1113, 1993. doi: 10.1109/34.244673. 36

[159] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 42

[160] R. Xie, C. Wang, and Y. Wang. Metafuse: A pre-trained fusion model for human pose estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13683–13692, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. doi: 10.1109/CVPR42600.2020.01370. URL `https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.01370`. 46

[161] Y. Xie, H. Jiang, G. Gkioxari, and J. Straub. Pixel-aligned recurrent queries for multi-view 3D object detection. In *ICCV*, 2023. 38

[162] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 28, 47, 48

[163] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 34

[164] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1385–1392, 2011. 41

[165] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua. LIFT: Learned invariant feature transform. In *European Conference on Computer Vision (ECCV)*, pages 467–483. Springer, 2016. 16

[166] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *Computer Vision – ECCV 2016*, pages 649–666. Springer International Publishing, 2016. 32

[167] R. Zhang, P. Isola, and A. A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017. 32

[168] S. Zhang, C. Wang, W. Dong, and B. Fan. A survey on depth ambiguity of 3d human pose estimation. *Applied Sciences*, 12(20), 2022. ISSN 2076-3417. doi: 10.3390/app122010591. URL https://www.mdpi.com/2076-3417/12/20/10591. 2

[169] S. Zhang, J. Wang, Y. Xu, N. Xue, C. Rupprecht, X. Zhou, Y. Shen, and G. Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21936–21947, June 2025. 52

[170] Z. Zhang, C. Wang, W. Qiu, W. Qin, and W. Zeng. Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild. *International Journal of Computer Vision*, 129:703 – 718, 2020. 46

[171] Y. Zhong, J. Yang, P. Zhang, C. Li, N. C. F. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li, and J. Gao. Regionclip: Region-based language-image pretraining. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16772–16782, 2021. 35

[172] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra. Detecting twenty-thousand classes using image-level supervision. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision – ECCV 2022*, pages 350–368, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20077-9. 35

Printed by Media-Tryck, Lund 2025 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

Doctoral Theses in Mathematical Sciences 2025:3
ISBN 978-91-8104-604-5
ISSN 1404-0034