

UNIVERSITÀ
di VERONA
Department
of COMPUTER SCIENCE

Deep Learning for Human Behaviour Understanding

A Comprehensive Study of Trajectory, Pose, and Gaze in Social and Human-Robot Interaction Scenarios

Andrea Toaiari

Supervisor(s):
Prof. Marco Cristani, Supervisor
Prof. Cigdem Beyan, Co-supervisor

Coordinator:
Prof. Ferdinando Cicalese

Department of Computer Science
University of Verona

Doctoral Dissertation
Doctoral Program in Computer Science
37^{th} *cycle, IINF-05/A*
2025

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License, Italy. To read a copy of the licence, visit the web page: http://creativecommons.org/licenses/by-nc-nd/3.0



Attribution —You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner but not in any way that suggests the licensor endorses you or your use.



NonCommercial —You may not use the material for commercial purposes.



NoDerivatives —If you remix, transform, or build upon the material, you may not distribute the modified material.

Deep Learning for Human Behaviour Understanding: A Comprehensive Study of Trajectory, Pose, and Gaze in Social and Human-Robot Interaction Scenarios

Andrea Toaiari

PhD Thesis, Verona, May 9, 2025

Acknowledgements

I would like to begin my potentially endless list of acknowledgements from the main person responsible for the existence of this thesis (right after me), Prof. Marco Cristani. My deepest gratitude goes to him for convincing me to pursue this journey, for offering me the opportunity to walk it with him and for his unconditional support during these years. Despite the difficulties and countless obstacles, in one way or another, he was always able to remind me why I decided to start this path. I would also like to thank Prof. Cigdem Beyan for her contribution in the last months of my doctoral program, for the face-to-face discussions on how to approach new projects, and for the pride she always took in our achievements. Besides them, my sincere thanks go to Prof. Alessandro Vinciarelli for hosting me at the Social AI CDT Lab at the University of Glasgow and for teaching me a different and novel way of approaching our craft. None of the work I have done would have been possible without the help of my co-authors; thanks to all of you.

This experience would not have been the same if I had not shared it with exceptional people. First of all, I would like to thank my friends and colleagues Andrea Avogaro, Luigi Capogrosso and Geri Skenderi, for helping me get to the end, for always being there and available for a chat and also for giving me the opportunity to cultivate together the next step in this adventure. My most sincere thanks also go to the other friends I made within the walls of our messy little laboratory: Federico Cunico, Federico Girella, Christian Joppi and Francesco Taioli. It has been great to share these years with you, discuss the newly released papers, and joke around during lunch breaks. My thanks also to all the other people with whom I shared the long days at the university, in Verona and in Glasgow, both the good and the bad ones. There are too many of you, and it would be too difficult to name you all, but my thanks definitely go to you as well.

Outside the university, I cannot help but thank my lifelong friends, those with whom I grew up and who have put up with me so far: Filippo Lovato, Valeria Lovato, and Caterina Roncari. You all have been vital, even though I may not have said it

enough. I also feel I have to thank the new friends who have come and those I no longer hear from; each has played their part, and I would not change anything that has happened to me. I then feel an obligation to thank my inspirations, the musicians who composed the music I listened to, the developers who created the video games I played, the writers who wrote the books I read, and the directors who directed the movies I watched.

Finally, I would like to express my greatest gratitude to the people who have filled my life, starting with Francesca, my partner and best friend, who has put up with me and supported me all these years, and it is probably mostly thanks to her that I have come this far. Then, thanks to my brother and my parents, my happy place, because when I am with them, I never feel alone, and life seems easier.

The best is yet to come.

Abstract

Human behaviour and communication are deeply rooted in non-verbal cues, including body movements (Kinesics), the spatial relationships between individuals (Proxemics) and the analysis of their focus of attention (where and what they are looking at). These subtle cues form an essential part of social interactions and the relationship a person has with their environment, and understanding them is critical for a range of applications, from Augmented Reality (AR) and Human-Robot Interaction (HRI) to surveillance and healthcare-related studies. With recent advances in deep learning and computer vision algorithms, methodologies for analysing such behaviours have become increasingly sophisticated, enabling researchers to model trajectories, poses and gaze with unprecedented accuracy. However, these tasks are complicated by multiple factors: the dynamic nature of human behaviour, the need for privacy-preserving methodologies, the challenge of modelling complex 3D environments and the lack of adequate datasets to explore dedicated use cases.

This thesis explores how deep neural networks can be leveraged to analyse human kinesics and proxemics, focusing on three key aspects: Human Trajectory Prediction (HTP), Human Pose Estimation (HPE) and Human Pose Forecasting (HPF), and the Gaze Estimation and Gaze Target Detection (GTD) tasks. The thesis is structured around a series of interrelated works that address both fundamental challenges in data collection and ethical concerns around privacy while advancing state-of-the-art methods for understanding human behaviour.

We present two datasets: the first one is related to the Human-Environment Interaction (HEI) scenario and designed to study the Visual Selective Attention (VSA) of a single and a couple of people moving in a laboratory, considering Social Signal Processing (SSP) insights and mapping the attention on a 3D scene model. The second one tackles HRI from the perspective of a Spot quadruped robot, and it's equipped with HPE, HPF, and Collision Prediction (CP) benchmarks on the video data acquired by the robot, which poses an interesting challenge given the often non-visible parts of the human body.

We also present a novel model for HTP in indoor environments, where it is crucial to cope with the different path options and countless layout configurations in contrast to the outdoor counterpart. This is achieved by combining equivariant and invariant geometric feature learning modules and a self-supervised vision representation extracted from the environment map. Regarding the GTD task, i.e., finding the target of the human gaze, we present a novel Multi-Task Learning (MTL) methodology for 3D GTD exploiting the upper-body pose and depth map to preserve the privacy of the observed subject and achieving state-of-the-art results on the most comprehensive dataset available in the literature. Finally, we propose a real-world application, where a Gaze Estimation algorithm working in real time has been integrated inside a smart shopping mall, tracking the users' favourite products and offering a personalized shopping experience for each customer.

In summary, this thesis analyses the application of deep learning and computer vision techniques to understanding and interpreting human behaviour when observed through the lens of a camera. We derive exciting insights, from the need for larger-scale datasets to tackle these tasks in the new foundation model-dominated landscape to the difficulty in solving these challenges in a world where privacy and safety are more critical than ever.

Table of Contents

List of Figures **xi**

List of Tables **xvii**

1 Introduction **1**
1.1 Human Behavior Understanding 1
1.1.1 High-level to Fine-grained Human Behaviour Analysis . . . 3
1.2 Thesis Overview 6
1.2.1 Contributions 7
1.2.2 Outline 8
1.2.3 Publications 10

2 Background **13**
2.1 Human Trajectory Prediction 14
2.2 Human Pose Estimation and Pose Forecasting 16
2.3 Gaze Analysis 23

3 Human Trajectory Prediction *in Indoor Environments* **27**
3.1 Introduction 29
3.2 Related Work 31
3.2.1 Indoor Human Trajectory Prediction 31
3.2.2 Equivariant and Invariant Graph Neural Networks 31
3.2.3 Self-Supervised Vision Representation 32
3.3 Method 33
3.3.1 The **SITUATE** Prediction Network 34
3.3.2 Feature Initialization 36
3.4 Evaluation 37
3.4.1 Evaluation Setup 37

3.4.2 Indoor Human Trajectory Prediction Results 39
3.4.3 Outdoor Human Trajectory Prediction Results 40
3.4.4 Ablation Study 41
3.5 Findings 42

4 Human Pose Estimation and Forecasting *from the Robot's Perspective* 43
4.1 Introduction 45
4.2 Related Work 47
4.3 The **HARPER** Dataset 49
4.3.1 Acquisition Setup 50
4.3.2 Actions and Annotations 51
4.3.3 Dataset Statistics 52
4.4 Evaluation 54
4.4.1 3D Human Pose Estimation 55
4.4.2 3D Human Pose Forecasting 56
4.4.3 Collision Prediction 58
4.5 Findings 60

5 Gaze Estimation and Gaze Target Detection *in the 3D Space* 61
5.1 Introduction 63
5.2 Related Work 65
5.2.1 Gaze Target Detection in 2D 65
5.2.2 Gaze Target Detection in 3D 66
5.2.3 Visual Attention Mapping 67
5.2.4 The Social Signal Processing Point of View 67
5.3 Visual Selective Attention Towards the 3D Scene 68
5.3.1 The **SCENE-pathy** Dataset 69
5.3.2 The Proposed Baselines 71
5.3.3 Evaluation 73
5.4 Privacy-Preserving 3D Gaze Target Detection 78
5.4.1 Methodology 79
5.4.2 Evaluation 83
5.5 Discovering the Gazed Target in a Shopping Mall 90
5.6 Findings 94

6 Conclusion **97**
6.1 Limitations and Future Works 100
6.2 Closing Remarks 103

References **107**

List of Figures

1.1 This thesis is organised into three interrelated macro-topics, each analysing an aspect of human behaviour addressed through novel computer vision and deep learning techniques. In addition, each topic is approached from an unconventional viewpoint w.r.t. other works in the literature, as shown in the *Scenarios* column. The last column summarises the contributions of this thesis, *e.g.*, innovative datasets enabling the study of these specific problems and solutions that consider the environment and the need for preserving the users' privacy and safety. 3

2.1 Examples of HTP applications. The first two images are outdoor predictions from the ETH [155] and UCY [117] datasets, while the third one is an example of indoor HTP result on the Supermarket dataset [66]. 14

2.2 Examples of 2D HPE predictions on the GFIE dataset [97]. 17

2.3 Examples of Gaze Estimation (first image) and GTD approaches (second and third images). The GTD examples are samples from the GFIE dataset [97]. 23

3.1 Examples of different trajectories from the Supermarket [66] dataset to show the difficulty of the indoor HTP task. In particular, the dataset showcases long trajectories (*Person 4*), self-loops (*Person 1* and *Person 3*), and confusing movements (*Person 2*) performed in an environment that strongly affects the people's paths. Specifically, the red circle represents the starting point of a trajectory, and the yellow star represents its final point. 29

3.2 In **SITUATE**, we first produce a feature vector regarding the scene using the self-supervised vision representation module. Then, a feature initialization layer is used to initialize geometric and pattern features. We then successively update the geometric and pattern features by the equivariant geometric feature learning and invariant pattern feature learning layers, obtaining expressive feature representation. We further use an invariant reasoning module to infer an interaction graph used in equivariant geometric feature learning. Finally, we use an equivariant output layer to obtain the final prediction. 35

4.1 **HARPER** showcase. (Top-left) We exploit the Spot on-board equipment to let the robot perceive people. (Top-right) Thanks to a 6-camera OptiTrack MoCap setup, we capture 3D human poses represented with 21 joints and 0.035 mm of error. (Second row) An additional external RGB camera shows the actions performed. (Third row) The gripper RGB camera PoV. (Fourth row) The gripper depth camera PoV, with the ground truth joints in yellow. 45

4.2 A 6-camera OptiTrack MoCap system covers a 6×6 squared meters area where users and Spot can freely move. The external RGB camera's FoV covers the setting. The 5 Spot on-body greyscale + depth cameras and the RGB-D frontal camera (gripper) cover the environment surrounding the robot. 50

4.3 Joints visibility from the robot's perspective. The left chart shows how many frames contain exactly n joints for $n = 1, \ldots, 21$. The right plot shows the percentage of frames in which the different parts of the skeleton are visible. 52

4.4 Distribution of distances between Spot and users (the distance considers the two closest joints of human and robot). Red columns correspond to distances lower than 10 cm, considered as cases of physical contact. 54

4.5 Results of 3D HPE from the robot's perspective. (a) On the left, the predicted 2D joints (in blue) by HRNet [181] and the corresponding ground truth joints (in red). On the right is the depth image with the same 2D detections. The depth will serve to do the lifting. (b) The lifted 3D poses alongside the complete OptiTrack skeletons. (c) MPJPE (in mm) for every visible joint (inside the depth FoV) on the test set. The size of the blobs is proportional to the errors, while colours are related to the number of times a joint is visible from the robot's perspective. 55
4.6 MPJPE for each joint using EqMotion with GT as input and a forecasting horizon of 1000 ms. 57
4.7 Qualitative results for the 3D HPF with the 1000 ms horizon. (a) shows the human pose forecasted in blue and the ground truth in red. At the end of the sequence, an accidental collision occurs. In (b), the collision (highlighted in green) is detected as explained in Section 4.4.3. The forecasting approach used is EqMotion [208] on the GT data. 58

5.1 (a) The 3D model of the **SCENE-pathy** showroom scenario, with the nine possible areas of interest highlighted by coloured bounding boxes. (b, c) On the left is an input video frame in which we highlight a single subject. On the right, the corresponding VSA estimation is in the form of a 3D map, where hotter colours indicate areas that attracted more VSA. 68
5.2 Qualitative results of an unsupervised single-subject experiment. (a) *Dynamic* module disabled. (b) *Dynamic* module active. Without the *Dynamic* module, the resulting VSA is definitely noisier. The target object was an advertising board, circled in green. 74
5.3 Qualitative results of a supervised experiment with a pair of subjects (S1 and S2) that had a dialogue in the center of the hallway. In (b) and (d), the objects circled in green correspond to the correct most attractive points for those VSA maps. (a) and (c) show instead the false positive targets (circled in red) obtained by disabling the *Dynamic* and *Social* modules. 76
5.4 (a) A frame from the GVEII dataset. (b) *SD-VSA* baseline. (c) *Vanilla* baseline: the *Dynamic* and *Social* modules are switched off. 76

5.5 We introduce a novel way to tackle the 3D Gaze Estimation and 3D GTD problems that, differently from the previous approaches, does not use the head crops of the observed people but rather employs upper-body skeletons, depth maps, and scene images where the head is blurred out. The new pipeline not only proved successful in tackling both tasks but offers a concrete way to preserve the identity of the involved people. The currently available state-of-the-art method already used depth maps (in magenta) in its pipeline, so we are not introducing additional information. 78

5.6 Overview of the proposed approach. (a) The 3D Gaze Estimation module predicts a gaze vector by exploiting the upper-body pose coordinates, processed by a simple MLP, and the convolutional features extracted from the depth map by a ResNet50 [87]. A single multi-head attention layer is applied to the concatenated features before the final MLP to regress the three vector coordinates. (b) The perception module converts the depth map to the unprojected point cloud, from which the 3D eye coordinates are subtracted, and computes two heatmaps highlighting the most interesting part of the scene. (c) The 2D gaze target heatmap is predicted by an encoder-decoder architecture operating on the concatenation of the scene image (with the blurred-out face), the head location mask and the two heatmaps. (d) During the evaluation, the most similar vector to the predicted vector $\hat{g}$ in the unprojected and translated point cloud is selected as the final gaze vector and the corresponding point in the original point cloud as the predicted 3D gaze target. 81

5.7 Qualitative results of our method on the GFIE dataset [97]. Each row represents a single sample. (First Column) We use resized depth maps as input for our Gaze Estimation module. (Second Column) The estimated upper-body skeletons are visualized on a black background before being normalized. (Third Column) Scene images where the faces are blurred out before being used by the encoder-decoder module to predict the 2D gaze target heatmap. (Fourth Column) The final point cloud of the scene, with the ground truth gaze vector in red and the estimated gaze vector in blue. . . . 88

5.8 Setup of the **I-MALL** demo. 91

5.9 The data aggregation module functional schema. 92

List of Tables

3.1 Deterministic HTP performance (ADE(m)/FDE(m)) on the THÖR and the Supermarket datasets. The **bold** and underlined fonts denote the **best** and second-best results. 39

3.2 Multi-prediction performance (ADE(m)/FDE(m)) on the THÖR and the Supermarket datasets. The **bold** and underlined fonts denote the **best** and second-best results. 39

3.3 Deterministic HTP performance (ADE(m)/FDE(m)) on the scenes of the ETH-UCY dataset. The **bold** and underlined fonts denote the **best** and second-best results. 40

3.4 Multi-prediction performance (ADE(m)/FDE(m)) on the scenes of the ETH-UCY dataset. The **bold** and underlined fonts denote the **best** and second-best results. 41

3.5 Ablation results (ADE(m)/FDE(m)) of **SITUATE**. We assess the contribution of the scene representation module and regularization methods in the deterministic prediction case. 41

4.1 Main HRI datasets revolving around human movement and its analysis. Values in the *Participants* column indicated with the asterisk (*) refer to datasets captured in uncontrolled scenarios. 48

4.2 **HARPER** actions. The expression *Contact* means that the distance between Spot and user is lower than 10 cm. 53

4.3 3D HPF errors. We provide the MPJPE expressed in mm with a prediction horizon of 400 and 1000 ms. The errors are computed for the particular frame for each action (first nine columns) as well as the average over all frames (*Average*) and the average over the last frame of each action instance (*Last Frame Average*). We use three different inputs: the ground truth OptiTrack skeletons (GT), the incomplete poses estimated by the 3D HPE baseline described in Section 4.4.1 or the filtered ground truth poses where only the joints seen by Spot were kept. *Depth* (D) means the poses are computed using the acquired depth maps. *Reconstruction* (R) means a trained diffusion-based imputation model reconstructs the poses before passing them to the forecasting method. 56
4.4 Performance of the different Collision Prediction methods with a 1000 ms horizon in terms of *Accuracy*, *Sensitivity*, and *Specificity* scores. The evaluation is divided into the four categories of contacts included in the **HARPER** dataset. 59

5.1 The accuracy in unsupervised experiments. 74
5.2 The accuracy in supervised experiments. In this case, the value tells us if we could identify the specific station assigned to subjects. . . . 76
5.3 Results on the GFIE dataset [97] for the 2D and the 3D scenarios. The results of the prior work are taken from [97]. *Dist*. stands for Distance. The **bold** numbers represent the **best** results. The second-bests are underlined. H, D, UP, and S stand for *Head crop*, *Depth map*, *Upper-body Pose*, and *Scene*, respectively. 87
5.4 Ablation study in terms of modality and the type of input scene image (*i.e.*, face blurred and not blurred) on the GFIE dataset [97]. *Dist*. stands for Distance. Note that depth maps and scene images are fundamental and must be included for the 3D GTD to function properly; thus, their removal cannot be ablated. 87

Chapter 1

Introduction

1.1 Human Behavior Understanding

Our world is undergoing a substantial transformation due to the increasing use of automated and intelligent technologies. Although it is important not to blindly rely on the use of such technologies, which are often fallacious from an ethical and individual independence standpoint, it is undeniable that their usefulness will prevail in many of the areas around us, from the way we shop to the way we help people in need. Among the places that will be most impacted, all those scenarios that require automatic analysis of human behaviour will incur a gradual and intensive change. From the beginning of Industry 5.0 [147], with worker-supportive technologies such as smart robots and advanced safety systems based on image analysis, to the development of software for autonomous driving and the use of new methodologies for healthcare, there are countless areas in which the analysis of human behaviour, observed from an image or a video, has and will grow in importance.

Since the birth of computer vision, analysing human behaviour from an image or a video captured by a camera has always interested researchers and industry practitioners. When observed from the outside, *i.e.*, by considering a third-person Point of View (PoV), human behaviour understanding consists of analysing the combination of the body movements, their relation with the environment and with other people and the person's future intentions. We can distinguish three main components of human non-verbal behaviour: kinesics, proxemics and gaze [85]. Kinesics, *i.e.*, the study of body movements and gestures, plays a central role in interpreting body language, where factors like posture, hand gestures, and facial expressions provide insight into a person's intentions and the activity they're performing. Proxemics ex-

amines the use of physical space in social and human-robot interactions, highlighting how interpersonal distances change based on cultural norms, relationship dynamics, context and the formation of groups. Meanwhile, the gaze represents a person's attention and focus, reflecting their interest, engagement, and future intentions. Together, these non-verbal behaviour elements, often encapsulated in the broad field of Social Signal Processing (SSP) [197], provide a nuanced understanding that extends beyond spoken words and is essential for developing perceptive Human-Robot Interaction (HRI) and Human-Environment Interaction (HEI) systems.

The rise of deep learning techniques, from early convolutional [115] and recurrent neural networks [92] to the recent advent of transformer architectures [195], diffusion-based models [91] and large foundation models [24], has resulted in solutions dedicated to analysing various aspects of human behaviour. The ability of these models to learn from large quantities of examples has offered performance previously not contemplated but has, on the other hand, imposed the need to acquire increasingly *large* and *specialized* amounts of data. *Large*, to cover the huge distribution of possibilities in the real world, and *specialized* in terms of annotations that must be sensor-accurate and enable learning different tasks. Predicting the trajectory of a person in a closed or open environment requires the annotations of its movements at a frame rate that is high enough to forecast the next couple of seconds. In contrast, the Human Pose Estimation (HPE) and Human Pose Forecasting (HPF) tasks require millimeter-accurate annotations for each human joint. While multiple sensors for acquiring this type of data exist and provide reliable measurements, their use is unsuitable when such perception systems are to be implemented on a large scale, limiting their use for the acquisition phase of a dataset or more specific use cases. In addition, sensors for acquiring body joints or head-mounted devices for gaze acquisition are often invasive, difficult to set up, non-scalable, and environmentally unfriendly. Computer vision-based tools make even more sense in this scenario, given that cameras are already installed in most public places and areas of interest and have a limited cost w.r.t. the number of considered subjects.

Despite the advantages described above, this can become a double-edged sword, increasing the risk of misuse of these technologies, harming the privacy of the subjects being analyzed and fostering a centralized control of information. For this reason, searching for methodologies that provide a helpful service while simultaneously preserving the users' privacy, such as exploiting features other than

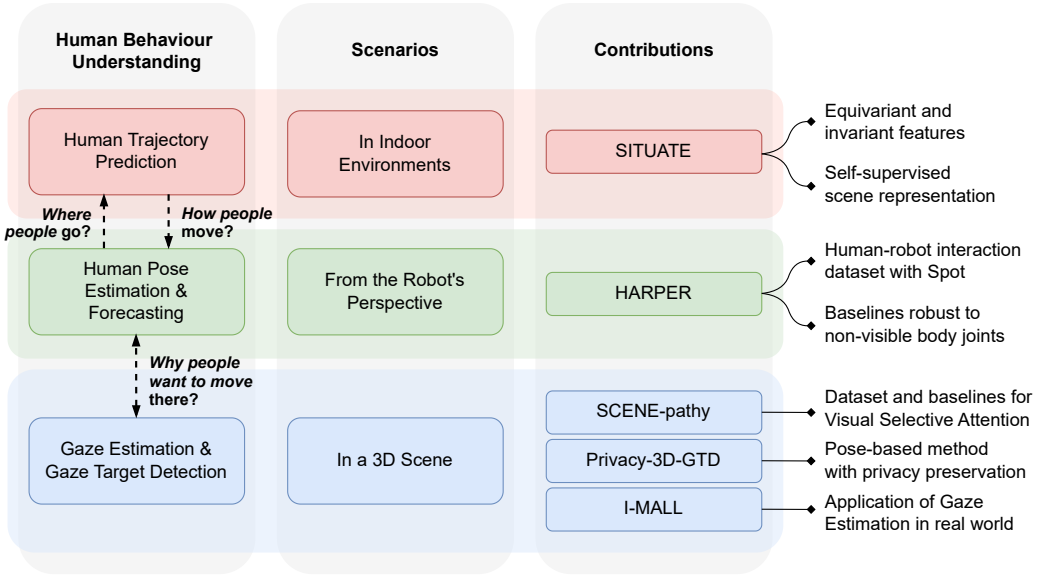


Fig. 1.1 This thesis is organised into three interrelated macro-topics, each analysing an aspect of human behaviour addressed through novel computer vision and deep learning techniques. In addition, each topic is approached from an unconventional viewpoint w.r.t. other works in the literature, as shown in the *Scenarios* column. The last column summarises the contributions of this thesis, *e.g.*, innovative datasets enabling the study of these specific problems and solutions that consider the environment and the need for preserving the users' privacy and safety.

appearance-related ones, is vital for the next generation of deep learning and computer vision-based applications.

1.1.1 High-level to Fine-grained Human Behaviour Analysis

Human behaviour can be interpreted at multiple levels of detail. From the most general, where we are interested in analysing the behaviour of multiple agents in a wide enough environment, to the most fine-grained, where we are interested in the movement of a person's individual limb or where they are directing their attention. This thesis focuses on three fundamental aspects of human behaviour related to the person's movements and intentions, and hierarchically ordered from the most general to the most specific: the movement trajectory [109], the body pose and motion evolution [100] and the gaze [160]. These three pillars are also interconnected because they can answer three crucial questions related to human behaviour analysis: *where* are people moving to (by forecasting their trajectories),

how they are moving there (by forecasting their body poses), and *why* are they moving there (by understanding their intentions and focus of attention).

Similar to how we perceive the world around us, each of these facets can be analysed by observing it through an image or video to predict its current state and possible future evolution. To answer the needs promptly, an automatic system must be able to not only predict the position and pose of a person at a certain time but also anticipate its future movements. The deep learning methods and datasets proposed in this thesis allow us to do just that. Not only pinpoint the location of a person, *e.g.*, by locating their body joints in the scene or their focus of interest in a specific frame, but to predict where and how they will move in subsequent instants of time, thus providing the opportunity for the intelligent systems to behave appropriately in response.

The following paragraphs present an overview of each topic we explored in this thesis and the particular research angle from which we tackled it:

Human Trajectory Prediction *in Indoor Environments*. Given an observed sequence of locations in space, Human Trajectory Prediction (HTP) consists of predicting the subsequent location the agents will occupy in the near future. Usually, this challenge is tackled in open spaces, outside, with few to no obstacles obstructing the people's movements. Differently, our work addresses the task of indoor HTP, an area that has been less explored compared to outdoor HTP, and specifically focuses on predicting indoor movements in restricted settings such as supermarkets or laboratories. While outdoor HTP has been widely studied, indoor environments add unique challenges due to structural elements like rooms, corridors, and obstacles that influence movement patterns. Indoor HTP requires understanding not just the path but also the layout and spatial semantics of the space. Understanding human movements in the next few seconds allows us to get a sense of what the person is doing, will be doing, and the interactions that might occur with other agents or the environment.

Human Pose Estimation and Forecasting *from the Robot's Perspective*. Estimating the pose of a human body requires locating each visible joint composing the human skeleton, where the structure and number of joints depend on the selected protocol. After estimating the position of each joint for an adequate number of consecutive frames, it is possible to adopt forecasting techniques to predict their

future movements, similar to what has been done for trajectories. Given that the joints' position provides more detailed information about the movement of a person, we can, in this case, make analyses or predictions about the activity performed by the observed subject. Moreover, by analysing the human pose, we can estimate the possible occurrence of interactions or collisions with other elements in the scene. HPE and HPF are usually performed using data acquired from a fixed PoV covering most of the scene, where all the human skeleton joints are visible. In this thesis, we instead analyse the case where estimation is done on visual inputs acquired from the sensors of a quadruped robot moving alongside the human. With the advent of Industry 5.0, more and more robots are being employed to support workers, with the need for seamless collaboration between them. The low height of a quadruped robot and the low quality of on-board video sensors present countless challenges. We propose a novel dataset to study this scenario, enriched with benchmarks for 3D HPE in the presence of multiple non-visible joints, 3D HPF, and robot-human Collision Prediction (CP). This framework allowed us to offer the first pipeline designed for collision avoidance and based on human poses from the perspective of a quadruped robot. This second pillar allows us to understand the pose and future movements of the observed people in more detail, understanding how they move and interact with their surroundings.

Gaze Estimation and Gaze Target Detection *in the 3D Space*. Gaze direction and focus of attention are fundamental cues to understanding people's intentions and what they are doing and will probably do. If a person is looking at an object and moving towards it, there is a good chance that they will interact with it. Two people looking at each other are probably having a social interaction, just as it is possible to detect the speaker in a group by observing the direction of the other interlocutors' gaze. Two main tasks can be addressed when analysing a person's gaze direction and focus: Gaze Estimation and Gaze Target Detection (GTD). Gaze Estimation [228], usually performed in 3D, consists of predicting the direction of the gaze, *i.e.*, the vector going from the eye's position to the region of the scene the person is looking at. It can be solved in multiple ways, by using the eye information if available (the person is front facing the camera and the resolution is high enough) or by adopting the head orientation to approximate the gaze direction. The second task is GTD [160], *i.e.*, finding the target of the gaze in the scene, be it a coordinate in pixel space or a 3D point in a point cloud.

While HTP and HPE could easily be addressed by a human operator just by looking at the input images, in this case, we are considering a non-trivial task requiring reasoning in three dimensions and leading to minor errors even through sensors. Unlike Gaze Estimation, the localization of the gaze target is performed in 2D in most cases, exploiting datasets where the annotations are manually performed, hence inaccurate by design. In this thesis, we propose two solutions designed to address different shortcomings related to gaze analysis through computer vision techniques. We first propose a dataset in which multiple people navigate alone or in pairs in a laboratory equipped with different industrial machines. At the same time, their gaze is captured using one of the most accurate sensors for gaze acquisition, the Hololens 2. The study is enriched with some dedicated baselines for estimating the Visual Selective Attention (VSA) towards the scene's elements, exploiting SSP insights, such as people's movement speed and the possible occurrence of social interactions. Contrary to the established GTD task, the attention was mapped continuously on a pre-built 3D point cloud of the scene by creating a temporal attention map, colouring even the laboratory's non-visible (from the third-person PoV camera) parts. In subsequent work, we tackle the traditional GTD task in the 3D space, providing a novel methodology capable of preserving the participants' privacy, using their upper-body pose as a proxy to estimate their gaze direction. Finally, we present an applicative scenario where we implement a Gaze Estimation pipeline to estimate the most interesting objects inside three shop windows, providing the mall's visitors with profiling of their interests to help them choose the most suitable products during their visit. This module is part of a project developed in collaboration with other research groups to create a smart shopping experience where consumers can enjoy services tailored to them using a combination of non-invasive deep learning and computer vision modules spread in the environment.

With these last works, we try to understand why a person is moving and acting in a certain way by analysing their gaze direction and focus of attention, which are usually good indicators when considering future intentions and activities.

1.2 Thesis Overview

This thesis aims to tackle various computer vision tasks for human behaviour understanding, such as HTP, HPE and HPF, Gaze Estimation and GTD, from the novel and original points of view not addressed by the previously available literature.

Furthermore, one of the key points is to provide exciting and practical solutions on an applicative industrial level by studying realistic problems and offering data collection for more specific and uncovered scenarios.

1.2.1 Contributions

In this thesis, we present our analysis of human behaviour through computer vision and deep learning techniques. The proposed solutions range from publishing new datasets enriched by dedicated benchmarks to creating methodologies to address specific use cases of well-known tasks, such as HTP in indoor environments or privacy-preserving GTD in a 3D scene. The datasets and methods analysed present quantitative and qualitative analyses to ensure the highest reproducibility. In summary, the contributions presented in this work are:

- **SITUATE** [31], a novel approach for indoor HTP based on equivariant and invariant geometric feature learning modules and a self-supervised vision representation extracted from the environment map. Our methodology achieves state-of-the-art results in indoor scenarios and competitive results in outdoor scenarios, showing that indoor forecasting models generalize better than outdoor-oriented ones.

- The **HARPER** dataset [6], the first HRI dataset acquired from the PoV of the Spot quadruped robot. It features 15 actions, of which 10 include physical contacts, performed by 17 participants. The data are captured with the onboard sensors of Spot and an external OptiTrack MoCap setup. The dataset is enriched with three vision-based perception benchmarks revolving around the scarce visibility of the human body from the robot's perspective: 3D HPE, 3D HPF with the reconstruction of the joints outside the Field of View (FoV), and CP.

- The **SCENE-pathy** dataset [186] for analysing the VSA of multiple people moving in a closed environment. The dataset consists of many experiments, captured in the form of video streams, Hololens 2 position and gaze data, and questionnaires, where one or a couple of people were asked to move freely in an industrial laboratory containing nine different machines (the interest points). We then developed some baselines to map the attention towards the scene elements directly on the 3D model of the scene, exploiting 3D HPE and Gaze

Estimation techniques. The baselines were enriched with SSP findings: the movement speed and the occurrence of social interaction should be considered when studying the focus of attention.

- A solution for gaze analysis from multiple shop windows integrated into **I-MALL** [18], an ICT hardware and software infrastructure that enables the management of services related to places such as shopping malls, showrooms, and conferences held in dedicated facilities. **I-MALL** offers a network of services that perform customer behaviour analysis through computer vision and provide personalized recommendations made available on digital signage terminals. This industrial application creates a profile for each new customer, ranking the items the client found most interesting during the visit.

- A novel methodology for the 2D and 3D GTD tasks [187], achieving state-of-the-art results on the most comprehensive dataset currently available for this kind of problems (GFIE [97]). In this work, we exploited the upper-body pose of the involved subjects and the available depth maps to estimate a coherent 3D gaze vector describing their gaze's direction. The human body pose can provide valuable information about the head pose, which approximates the gaze direction and the position of the arms and hands linked to the activity the person is performing and the objects they are likely focusing on. By blurring the faces of the subjects, we also ensured the preservation of their privacy, often overlooked in most computer vision methodologies for the automatic analysis of human behaviour, where the face appearance is used to extract multiple features. The Gaze Estimation module we developed is then integrated into a pipeline for solving the 2D and 3D GTD tasks, *i.e.*, finding the looked-at point in the input image or point cloud, respectively.

1.2.2 Outline

This thesis is structured as follows.

Chapter 2 provides an extensive review of the literature on human behaviour understanding, focusing on three major aspects investigating how humans move and interact with their environment. Section 2.1 presents the task of HTP, *i.e.*, predicting the movement trajectory of a person given some previously observed steps. In Section 2.2, we introduce the task of HPE, *i.e.*, finding the location of the human skeleton joints by analysing images or videos in both 2D and 3D scenarios.

Following that, we introduce a related task, HPF, *i.e.*, predicting the future locations of the observed human joints. Section 2.3 illustrates the Gaze Estimation task, *i.e.*, finding the gaze direction of one or multiple people by using an image or video of the scene as input. This chapter lays the groundwork for the approaches developed in subsequent chapters by identifying critical research gaps and motivating the need for innovative solutions.

In Chapter 3, the focus is on the task of HTP, particularly in complex indoor environments. Unlike outdoor trajectory prediction, which assumes open spaces, indoor scenarios require models to account for structural constraints like walls, obstacles, and varying layouts. The chapter introduces the **SITUATE** architecture, which leverages a combination of equivariant and invariant geometric feature learning modules with self-supervised vision representations. Section 3.2 provides a background of HTP in indoor environments, while Section 3.3 describes our methodology for solving it. Comprehensive experiments on real-world datasets in Section 3.4 demonstrate the model's capability to outperform existing methods, particularly in handling diverse path options and unpredictable movement patterns. The findings are summarized in Section 3.5.

Chapter 4 addresses the challenge of estimating and forecasting human poses when viewed from a moving quadruped robot, in our case, the Spot by Boston Dynamics. This scenario presents unique difficulties, including limited FoV, partial occlusions, and the low height of the robot's cameras. To tackle these challenges, we introduce the **HARPER** dataset in Section 4.3, providing annotated 3D human poses and human behaviour benchmarks specifically designed for HRI scenarios. Section 4.2 gives an overview of the HRI datasets in literature and their differences when compared to ours. The chapter proposes novel benchmarks for 3D HPE (in Section 4.4.1), 3D HPF (in Section 4.4.2) and CP (in Section 4.4.3), incorporating robust handling of non-visible joints and reconstruction of the missing ones. Experimental results discussed in Section 4.5 validate the effectiveness of the proposed techniques, paving the way for more seamless collaboration between humans and robots in industrial and social settings.

Chapter 5 focuses on the Gaze Estimation and GTD tasks in 3D environments, proposing a novel dataset for the study of VSA, dubbed **SCENE-pathy**, in Section 5.3, a privacy-preserving methodology that uses upper-body skeletons and depth maps instead of traditional head-crop-based approaches in Section 5.4, and a real-world Gaze Estimation application in Section 5.5. Understanding gaze behaviour is

critical for interpreting human intentions and actions. First, we give an overview of the literature concerning GTD in both 2D and 3D scenarios, the mapping of visual attention, and the SSP standpoint in Section 5.2. In Section 5.3.1, Section 5.3.2 and Section 5.3.3 we present the **SCENE-pathy** dataset, the SSP-based baselines to study VSA and the evaluation we performed on the new data, respectively. Section 5.4.1 introduces the privacy-preserving pipeline we adopted to solve the 2D and 3D GTD tasks. Experimental evaluations in Section 5.4.2 demonstrate that the proposed approach achieves state-of-the-art performance while preserving individual privacy. Section 5.5 discusses a real-world implementation of these methodologies in a smart shopping mall, where gaze data are used to personalize the customer experience. Section 5.6 summarizes the works presented in the chapter, discussing the particular limitations of the proposed methods.

Finally, Chapter 6 synthesizes the contributions of the thesis, reflecting on the implications of the proposed methodologies for human behaviour understanding. Limitations and future research directions are discussed in Section 6.1, which include the challenges of scaling the approaches to large, diverse datasets and the need for more precise annotations. The use of large foundation models on the one hand and the need for fast, on-the-edge models for real-time computations on the other are both attractive options for novel and more human-centered applications.

1.2.3 Publications

The publications carried out during the author's PhD program are listed below in chronological order. Please note that some of these articles are not part of this thesis. Those included are highlighted in **bold**.

- Cunico, F., Toaiari, A., and Cristani, M. (2022).
  A Masked Face Classification Benchmark on Low-Resolution Surveillance Images.
  In International Conference on Pattern Recognition (pp. 49-63).

- Becattini, F., Becchi, G., Ferracani, A., Bimbo, A.D., Presti, L.L., Mazzola, G., Cascia, M.L., Cunico, F., Toaiari, A., Cristani, M., and Greco, A. (2022).
  I-MALL An Effective Framework for Personalized Visits. Improving the Customer Experience in Stores.
  In Proceedings of the 1st Workshop on Multimedia Computing Towards Fashion Recommendation (pp. 11-19).

- Toaiari, A., Cunico, F., Taioli, F., Caputo, A., Menegaz, G., Giachetti, A., Farinella, G.M., and Cristani, M. (2023).
  SCENE-pathy: Capturing the Visual Selective Attention of People Towards Scene Elements.
  In International Conference on Image Analysis and Processing (pp. 352-363).

- Skenderi, G., Capogrosso, L., Toaiari, A., Denitto, M., Fummi, F., Melzi, S., and Cristani, M. (2023).
  Disentangled Latent Spaces Facilitate Data-Driven Auxiliary Learning.
  arXiv preprint arXiv:2310.09278.

- Avogaro, A., Toaiari, A., Cunico, F., Xu, X., Dafas, H., Vinciarelli, A., Li, E., and Cristani, M. (2024).
  Exploring 3D Human Pose Estimation and Forecasting from the Robot's Perspective: The HARPER Dataset.
  In 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).

- Capogrosso, L., Toaiari, A., Avogaro, A., Khan, U., Jivoji, A., Fummi, F., and Cristani, M. (2024).
  SITUATE: Indoor Human Trajectory Prediction through Geometric Features and Self-Supervised Vision Representation.
  In International Conference on Pattern Recognition.

- Toaiari, A., Murino, V., Cristani, M., and Beyan, C. (2024).
  Upper-Body Pose-based Gaze Estimation for Privacy-Preserving 3D Gaze Target Detection.
  In Computer Vision – ECCV 2024 Workshops.

Chapter 2

Background

This chapter presents an overview of the literature and previous works related to the main topics covered by this thesis. The topics are introduced here in general terms, only to be further elaborated in the following chapters with respect to the specific application scenarios that we selected for investigation. In Section 2.1, we introduce the task of Human Trajectory Prediction (HTP), which we later considered in its indoor variant. Subsequently, in Section 2.2, we describe methodologies and datasets to address the Human Pose Estimation (HPE) and Human Pose Forecasting (HPF) tasks, focusing on the case where the input is an image or a video. Finally, we cover the literature related to human gaze analysis, addressing the task of Gaze Estimation in Section 2.3, again from a vision-based perspective. Although we are more interested in the specific subtask of Gaze Target Detection (GTD), this broader review allows us to better frame our contribution within the larger context of gaze analysis research.

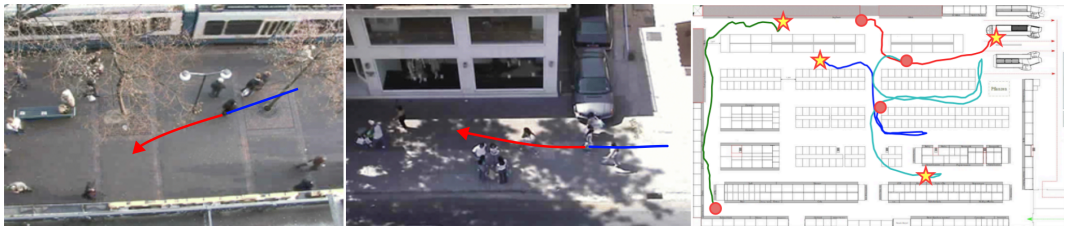


Fig. 2.1 Examples of HTP applications. The first two images are outdoor predictions from the ETH [155] and UCY [117] datasets, while the third one is an example of indoor HTP result on the Supermarket dataset [66].

2.1 Human Trajectory Prediction

Predicting the evolution of a pedestrian trajectory in the future is a long-standing task whose interest is constantly renewed by the emergence of new scenarios that can benefit from it, *e.g.*, autonomous driving [167]. Other application domains include Human-Robot Interaction (HRI) scenarios, where robots can plan their movements based on the predicted human (or different agents) trajectories, crowd-control systems and advanced commercial services. Given a sequence of observed positions, typically represented as spatial coordinates in a 2D or 3D space, Human Trajectory Prediction (HTP) aims to predict the subsequent trajectory the observed person will likely follow, composed of a certain number of future steps. Some examples of HTP are shown in Figure 2.1. When proposing a methodology to tackle HTP, one should take care of several aspects, from the environment's geometry [134] to the presence of obstacles [83] or different contextual cues and the possible interactions between multiple agents [7]. Some traditional methods to approach this task involved force models [88], Markov models [109], and Recurrent Neural Networks (RNNs) [223]. Notably, considering common sense rules and conventions that humans observe in social spaces helps to manage simultaneous predictions in crowded scenes [3]. A lot of deep learning-based methodologies have been applied successfully to forecast pedestrian trajectories, such as Graph Neural Networks (GNNs) [216, 8, 174], generative models [80, 51, 122], transformers [72] and Conditional Variational AutoEncoders (CVAEs) [133, 210]. More recently, diffusion models have also been applied to solve this problem [75, 136].

Deep Learning-based Approaches for HTP. Recent and well-known methodologies for HTP include Trajectron++ [168], TransF [72], MemoNet [207], EqMotion [208], LED [136], SocialCircle [205] and SingularTrajectory [10].

Trajectron++ [168] is a generative graph-based multi-agent approach built for robotics use cases. It incorporates heterogeneous data the agents provide and can account for their dynamics. [72] applied for the first time a transformer network to solve this task by using simple models where each person is modelled separately, without interactions with other agents or the environment. [136] leveraged a trainable leapfrog initializer to directly learn an expressive multimodal distribution of future trajectories, drastically reducing inference time. SocialCircle [205] and Singular-Trajectory [10] are two of the most recent methods for HTP. The first one [205] analysed each person's surroundings by utilizing a novel angle-based trainable social interaction representation inspired by the echo mechanism used by water animals. The second one [10] proposed a diffusion-based HTP framework to reduce the performance gap across five different tasks: deterministic, stochastic, domain adaptation, momentary observation, and few-shot trajectory prediction. This was achieved by building a custom embedding space where different motion patterns can be projected and refined by a cascaded denoising process, generalizing across modalities and trajectory lengths.

Datasets for HTP. Multiple datasets have been introduced to benchmark the HTP task, providing ground truth annotations ranging from basic movement trajectories to more fine-grained information, including agents' orientation, obstacle locations, group formations, and activity labels. Most datasets are based on video data acquired from a top-down view using camera sensors and enriched with manual annotation. In many cases, automatic interpolation is applied to fill in the frames between annotated positions. Commonly used outdoor HTP include the ETH [155] and UCY [117] dataset group, the Stanford Drone Dataset [164], which also contains cyclists and vehicles trajectories, the Edinburgh dataset [132], notable for its large number of tracks (92,000) compared to the competitors but presenting a single scene, the KITTI dataset [68], acquired using a Velodyne sensor and four cameras mounted on a car and the Oxford Town Center Dataset [19], which depicts a shopping street with diverse pedestrian activities and show windows.

In contrast, there is a notable lack of comprehensive indoor datasets, despite the fact that indoor trajectories often exhibit more complex and non-linear patterns [165] compared to their outdoor counterparts. When traversing indoors, our immediate movement decision is influenced by the objects in our path and the surrounding walls [134]. By comparison, outdoor environments typically offer broader movement

areas with fewer spatial constraints and limited contextual cues, allowing subjects to spread out more freely. Examples of indoor datasets include the Supermarket dataset [66], which captures customer trajectories in a retail environment, and the THÖR dataset [166], featuring a HRI context where multiple people with different roles (and activities) move in a constrained environment alongside a service robot. This scarcity of indoor datasets, combined with their inherent complexity, has led to HTP methodologies that often exhibit limited generalization capabilities when transferred from outdoor to indoor environments.

Our Contribution

In Chapter 3, we present a methodology for analyzing human trajectories in indoor environments, including settings such as a supermarket and a laboratory for HRI experiments. Specifically, by extracting a semantic representation of the scene layout and leveraging both equivariant and invariant features, the proposed approach achieves state-of-the-art performance in indoor environments while maintaining competitive results in widely used outdoor benchmarks. This represents the first and most general level of understanding human behaviour through computer vision methods. Predicting the future trajectory of one or more individuals enables us to answer the fundamental question: *Where* will this person move in relation to nearby agents (such as people and robots) and the surrounding environment?

2.2 Human Pose Estimation and Pose Forecasting

Human Pose Estimation

Vision-based Human Pose Estimation (HPE) refers to the automatic localization of human joints (or keypoints) in images or videos [234]. Each keypoint represents a 2D or 3D coordinate associated with a specific joint (*e.g.*, left shoulder, right elbow, right ankle, etc.), while valid connections between joints are referred to as edges or limbs (Figure 2.2). This task presents several challenges, including variability in input resolution (and the corresponding size of the people), occlusion, poor lighting conditions, varying clothing (*e.g.*, dark garments), and the demand for real-time processing.

Accurate estimation of human pose has proven valuable across multiple domains. It supports augmented and virtual reality applications, enables motion capture for film

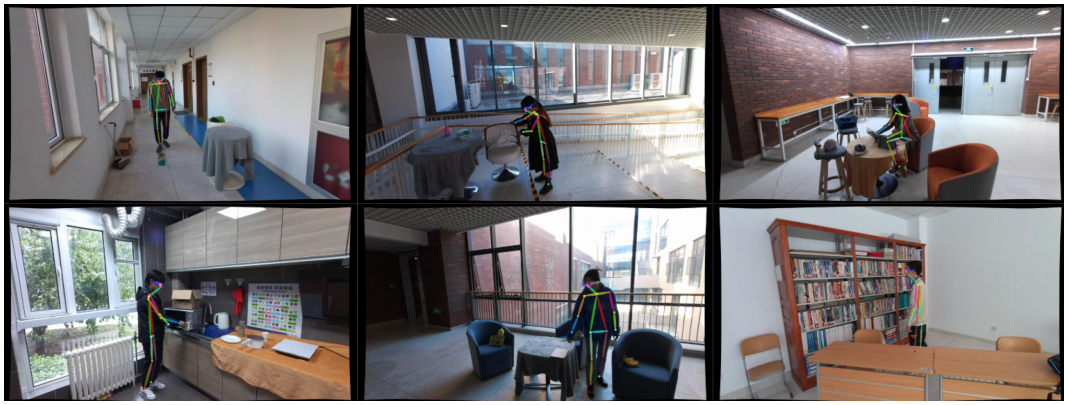


Fig. 2.2 Examples of 2D HPE predictions on the GFIE dataset [97].

and gaming by animating 3D characters, and enhances industrial HRI scenarios by allowing robots to interpret human collaborators' movements in shared environments. Healthcare also benefits through assistive technologies that can detect harmful or abnormal movements, and the sports industry increasingly leverages pose estimation for performance analysis and injury prevention.

Furthermore, HPE is a fundamental component in the motion analysis pipeline, typically serving as a first step to future pose prediction (discussed in the next section). While HPE in 2D space has reached outstanding results, estimating the 3D location of each joint from a single image or video frame remains a particularly difficult task, given the ambiguity in estimating the body's depth. Multi-view 3D HPE can alleviate this ambiguity while introducing additional complexities that continue to drive research in this area.

Deep Learning-based 2D HPE. Single-person HPE assumes the input image contains only one subject, and it represents the simplest case. Modern techniques are divided into two primary categories, based on how they solve the task: regression-based and heatmap-based methods. Regression-based methods directly map input images to the 2D coordinates of body joints in an end-to-end fashion. Methods such as DeepPose [190] employed cascaded deep neural networks to refine pose predictions iteratively. Advances such as Compositional Pose Regression [182] introduced bone-based representations, capturing structural dependencies between joints and improving estimation accuracy.

Heatmap-based methods, on the other hand, predict a probability distribution over the image space for each keypoint, represented as heatmaps. Techniques like

Convolutional Pose Machines [204] and Stacked Hourglass Networks [150] refine these heatmaps iteratively, achieving superior accuracy by leveraging spatial and contextual information. For images containing multiple people, HPE becomes more challenging due to the need to localize individuals and associate keypoints to specific human subjects. The two predominant approaches to address this task are top-down and bottom-up. Top-down methods first detect individuals using a person detector (*e.g.*, the widely knows Yolo [162] detection algorithm), then apply single-person pose estimators to each detected region. Some examples are HRNet [181], PPT [128], ViTPose [211] and PCT [70]. While accurate, their performance scales poorly with the number of individuals in the scene. Instead, bottom-up methods simultaneously detect all keypoints in the image and then group them into individual poses. Some examples of this kind of algorithm are OpenPose [29], HigherHRNet [38], and DEKR [69]. Although computationally efficient, bottom-up methods face challenges in correctly associating keypoints to specific people, especially in crowded scenes.

Generally, for 2D HPE, it is essential to consider partial and occluded poses, which means learning delicate local representations of body parts. For instance, in [28], the authors obtained a local representation that retains rich spatial information to perform accurate keypoint detection. The authors of [29] created a bottom-up framework that is very effective with heavily occluded environments and offers real-time performance, thanks to the use of Part Affinity Fields for body parts localization. Similarly, in [158], a two-stage pipeline has been proposed to estimate human pose under strong occlusions. HRNet [181] is among the most widely used approaches and differs from other methodologies in maintaining high-resolution representations throughout the whole process, combining them in parallel with high to low-resolution sub-networks. This architecture does not use intermediate heatmap supervision, and heatmaps are regressed using a Mean Squared Error (MSE) loss. The work presented in [211] used a plain transformer network, achieving noteworthy results with a compact model size.

Among the several use cases in which HPE approaches can be applied, one of the most challenging scenarios is undoubtedly the one dealing with small or low-resolution people, which often occurs in applications such as video surveillance of crowds or autonomous driving. Tiny People Pose [149] proposed a specific solution for this problem, *i.e.*, predicting a posterior probability over pose configurations. In particular, the authors defined a neural network architecture that explicitly expresses uncertainty, training it by explicitly minimizing a novel loss function based on the

data log-likelihood. They also estimated the posterior probability maps for all joints as a semi-dense sub-pixel Gaussian mixture model. In addition to single-frame estimation, several methods have explored leveraging temporal information from video sequences to improve prediction consistency and robustness, as demonstrated by [124]. The additional information can help maintain consistency but poses a non-trivial constraint in the type of data a network can consume as input.

The most widely used datasets for 2D HPE include the MPII Human Pose Dataset, which contains over 40,000 manually annotated human poses covering a broad range of activities. The COCO dataset is one of the largest available, with more than 200,000 annotated individuals and multiple configurations tailored for different pose estimation benchmarks. The PoseTrack dataset focuses on video-based HPE and provides over 150,000 annotations across various video sequences. Lastly, the CrowdPose dataset offers approximately 80,000 annotated human poses, combining samples from earlier datasets and emphasizing crowded scene scenarios.

Deep Learning-based 3D HPE. Extending HPE into the 3D space involves predicting the depth of body joints in addition to their 2D locations. This task is inherently more complex due to depth ambiguities, occlusions, and the ill-posed nature of mapping 2D projections back into 3D space. Single-person 3D HPE methods can be broadly categorized based on their approach to representation. Direct estimation methods infer 3D coordinates directly from the input image without intermediate representations. Techniques like these often employ volumetric representations or regression frameworks to predict joint coordinates in 3D space [118]. Leveraging the success of 2D HPE, 2D-to-3D lifting [235, 35, 209, 140, 232] approaches predict 2D keypoints and then map them into 3D space using lifting models. Graph Convolutional Networks (GCNs) [42] have successfully captured spatial dependencies between joints in this context.

For 3D HPE, learning the representation of human motion is critical. In [236], the authors proposed a unified 3D HPE framework relying on a dual-stream spatiotemporal transformer and directly analysing the video input. In [94], lifting was performed with a GCN, starting from the 2D pose sequence. Similarly, [222] lifted the 2D pose sequences thanks to a mixed spatiotemporal transformer-based model. The authors of [183] created a similar method for the lifting with a transformer-based architecture, relying on a criss-cross attention block. However, diffusion models have shown their potential in solving various computer vision tasks, including 3D

HPE. For instance, [173] performed 3D HPE by sampling from pure 3D noise and performing a multi-hypothesis aggregation.

Unlike the numerous datasets available for 2D HPE, where data collection is primarily limited by time constraints, the acquisition and annotation challenges associated with 3D human poses have resulted in a relative scarcity of 3D pose datasets. Manual annotation of 3D keypoints is typically unfeasible, and the reliance on specialized sensor infrastructure further limits data diversity. As a result, most 3D pose datasets feature simplified and constrained scenes, often captured in controlled laboratory environments. This is demonstrated in the most used 3D pose dataset, the marker-based MoCap dataset Human3.6M [100], which presents 3.6 million 3D human poses in the format of single-person indoor sequences. Other commonly adopted datasets include 3DPW [137], MPI-INF-3DHP [141], HumanEva-I [177], TotalCapture [192] and the CMU Panoptic Dataset [103].

HPE with Other Sensors. In addition to images and videos, data from other sensors can also be used to perform HPE. This is the case, for example, in robotics, where it is possible to take advantage of depth sensors, usually RGB-D cameras, mounted on most of the available robots to overcome the ambiguity mentioned above in estimating 3D human poses. Considering the HRI field, the 3D HPE from the robot's Point of View (PoV) is a fascinating research line, and the approaches vary depending on the nature of the input data. Besides the narrow Field of View (FoV) of such RGB-D devices, occlusions significantly intensify the challenge [158].

For instance, in [67], the authors equipped a robot with an RGB camera and performed HPE and body part triangulation to make it move close to a human for social interaction. In [101], the authors proposed a method to inspect the human pose for physiotherapy tasks. They moved the robot around the human to find the optimal view angle for assessing the correctness of the performed physical exercise. [238] used an RGB-D-based approach to estimate the pose of a human in order to perform markerless action learning from human demonstration using a grasping robot. It is also worth mentioning that depth-only approaches exist, relying entirely on the 3D structure perceived by the depth sensors. The authors of [34] used raw depth maps obtained from a sensor mounted on a moving robot to get the optimal viewpoint for observing the human pose. This is similar to [101], but using only depth maps. In [138], the authors claimed that depth is enough to extract the human pose, proposing a multi-person HPE framework. Finally, it is worth

noting that LiDAR sensors have recently begun to be used as depth-only devices for HPE [220]. Other sensing modalities include marker-based systems, which offer high accuracy by minimizing the joint localization error, even though at the cost of reduced generalization to in-the-wild conditions due to their sensitivity to occlusion and constrained deployment settings. Additionally, radio frequency-based devices, such as WiFi transmitters, and Inertial Measurement Units (IMUs) have been employed either as standalone sensors or in conjunction with others for 3D HPE. These solutions offer enhanced privacy by avoiding using the appearance of subjects while still enabling effective pose estimation.

Human Pose Forecasting

In the previous section, we discussed the vision-based HPE task, *i.e.*, estimating a human pose given an image or video of the person performing it. A natural and fundamental question arises when considering a sequence of poses that describes a person's movement over time from a skeleton-based perspective: how will this movement evolve in the near future? This a fundamental computer vision research question, with applications extending beyond those already associated with HPE. In particular, predicting the evolution of human poses plays a key role in understanding performed activities, enabling proactive responses in domains such as HRI, surveillance, autonomous driving, sports analytics, and rehabilitation. Human Pose Forecasting (HPF), or Human Motion Prediction, addresses this challenge by forecasting a sequence of future poses for an observed individual based on a set of past pose observations.

While traditional methods exploited Markov models [116] and Gaussian Process dynamical models [199], from the introduction of deep learning models, this task follows two distinct lines of research. The first one focuses on the deterministic estimation of future poses [135, 78, 208], assuming that given a series of observed movements, there is only one possible sequence of future poses. This class of methods aims at predicting future poses in timespans of usually one second, making use of GCNs [129], simple feed-forward networks [78], the attention mechanism to investigate spatiotemporal representations [2], or invariant and equivariant features [208].

Stochastic methods [217], on the other hand, allow the estimation of multiple possible futures, sampling from a probability distribution based on past movements. In this category, the most commonly used methodologies include Generative Ad-

versarial Networks (GANs) [15] and Variational AutoEncoders (VAEs) [212]. To overcome the difficulties of these models in efficiently generating more diverse predictions, [217] proposed a novel sampling method to balance sample diversity and likelihood, using a diversity-promoting prior. In [184], the authors applied a conditional score-based diffusion model to the problem of time series imputation, *i.e.*, estimating the distribution of missing values from some observed data. This can be particularly useful in the case of missing, unstable, and noisy joints. Recently, [14] proposed to sample from a disentangled behavioural latent space to produce more realistic and coherent future poses. TransFusion [185] proposed a lightweight diffusion model which employs the discrete cosine transform to model motion sequences in the frequency space.

The most popular datasets for HPF are often MoCap datasets with multiple participants performing different everyday actions. Due to the practical advantages of recording 3D HPE datasets in video format, rather than capturing isolated images for each pose, these datasets have also gained popularity for addressing the 3D HPF task. Some examples of well-known HPF datasets are Human3.6M [100], its recent whole-body version H3WB [237], the in-the-wild 3DPW [137], and the AMASS dataset collection [131].

While this and the previous sections have focused primarily on skeleton-based HPE and HPF, it is important to note that deep learning offers a diverse range of approaches for analyzing and predicting human movement. A clear example is the whole-body mesh prediction research track, which extends beyond skeletal joints to estimate detailed body surface deformation, enabling more expressive applications in animation, simulation, and human-cloth interaction. Human Mesh Recovery (Human Mesh Recovery) methods [110, 46, 233] incorporate parametric body models, such as SMPL [125], to reconstruct detailed 3D meshes of the human body. They provide richer representations by including shape and texture information alongside pose estimation.

Our Contribution

Most of the datasets we reviewed for 3D HPE and 3D HPF are acquired from static viewpoints that fully capture the whole human body and are mainly used to simulate the scenario in which the camera is fixed on a wall or on a stationary robot. We decided to explore the opposite direction, investigating the setup in which RGB and depth cameras are mounted on a highly agile quadruped robot,

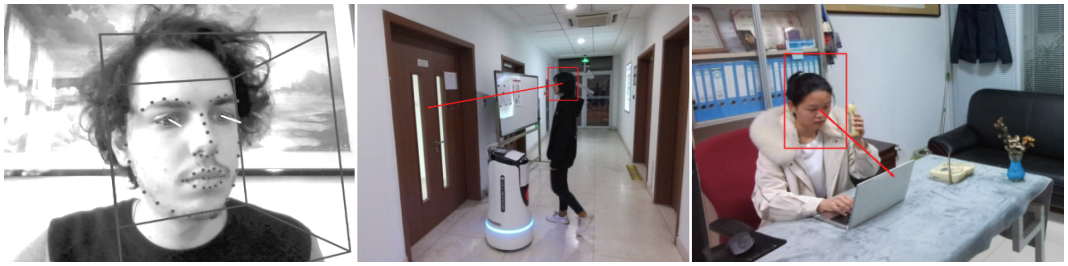


Fig. 2.3 Examples of Gaze Estimation (first image) and Gaze Target Detection (GTD) approaches (second and third images). The GTD examples are samples from the GFIE dataset [97].

moving in the same area as a human operator. In Chapter 4, we present a novel dataset featuring 17 participants, offering a double PoV (from the robot and from the panoptic MoCap system around the walking area) and a pipeline for tackling the 3D HPE, the 3D HPF and ultimately the Collision Prediction (CP) tasks from the quadruped robot's perspective.

In Section 2.1, we described HTP, a task whose goal is to understand *where* a person will move in space. Instead, estimating a person's pose using HPE techniques and its evolution over the next few seconds thanks to HPF methodologies allows us to answer the question: *How* will the person move their body from its current location to its future location?

2.3 Gaze Analysis

Computer vision-based gaze analysis comprises all those techniques that allow the study of eye movements, visual attention, and the resulting intent of multiple people observed using a camera or specialized gaze tracking sensors [71]. Gaze Estimation, in particular, involves estimating a person's gaze direction, usually in 3D, by exploiting facial and eye features, from hand-crafted descriptors to learned deep representations. It constitutes the basic task when it comes to gaze analysis, with applications in multiple fields, from Human Computer Interface (HCI) to healthcare and driving assistance systems.

While this thesis primarily focuses on the GTD subtask (described in Section 5.2), which requires to localize the target of the estimated gaze vector, it is important to first provide an overview of the Gaze Estimation task and its evolution in the past few years. Indeed, within the field of computer vision, Gaze Estimation has a

longer and more established research history than GTD, serving as the basis for many subsequent developments in gaze-related analysis. Some examples of both tasks are visible in Figure 2.3. Earlier work in Gaze Estimation often considered simple scenarios with carefully manually cropped head images as inputs. However, recent advancements, particularly with the adoption of deep learning methods, have shifted the focus to more challenging in-the-wild scenes. Subjects are often farther from the camera in these scenarios, resulting in lower-resolution facial images. There are two main approaches to Gaze Estimation: geometric eye models and appearance models. The first one focuses on leveraging ocular physical components such as the pupil, the cornea and the eyeball. While these model-based Gaze Estimation methods can be considered more precise, they usually require calibration steps that render them less efficient compared to the appearance-based ones. The advent of deep learning alleviated the need for complex parameter estimation in geometric-based approaches, while significantly advancing appearance-based methods by enabling richer visual representations. Studies have shown that using full facial embeddings extracted by pretrained networks, rather than focusing solely on eye images, can lead to better performance. This approach not only considers head pose but also leverages additional facial features to enhance accuracy. The most recent appearance-based models have shown increased interest in Convolutional Neural Network (CNN)-based solutions. For instance, [230] used both eye and facial features, while [112] developed a multi-channel network using eye images, full-face images, and face grid information. [37] introduced a model integrating statistical information within a CNN for eye images, utilizing dilated convolutions to retain high-level image features without reducing spatial resolution. Their GEDDNet model employs gaze decomposition with dilated convolutions for enhanced performance.

[60] enhanced gaze prediction accuracy by incorporating the head pose vector with VGG-extracted features from eye crops and using an ensemble scheme. [39] proposed CA-Net, which first predicts primary gaze angles from a face image and refines them with residuals from eye crops, using a bi-gram model to integrate primary gaze and eye residuals. [107] employed an LSTM-based temporal model with 7-frame sequences and adapted pinball loss for joint gaze direction and error bounds regression to improve accuracy.

More recently, transformers with self-attention modules have been applied to Gaze Estimation. One approach [152] used a transformer to extract crucial gaze features from images with high variance, filtering irrelevant information using con-

volution projection and preserving detailed features with a deconvolution layer. AGE-Net [114] proposed using two parallel networks for each eye image: one generating a feature vector with a CNN and the other generating a weight feature vector with an attention-based network. A related example in the 2D space can be found in [54], where a bi-dimensional gaze vector is estimated from the face landmarks.

There are multiple available datasets for Gaze Estimation, with various sizes, annotation precision, and diversity in terms of the number of subjects and environmental conditions. Some examples of datasets aimed at studying the appearance-based Gaze Estimation task are the MPIIGaze dataset [228], which contains more than 200,000 images collected from 15 participants during natural everyday laptop use, and the RT-GENE dataset [60], which also contains images of people wearing eye-tracking glasses. Two of the most widely used large-scale datasets are Gaze360 [107] and ETH-XGaze [226]. In particular, Gaze360 contains recordings from 238 subjects captured in both indoor and outdoor environments, offering labelled 3D gaze annotations across a wide range of head poses and distances.

A key observation is that many of these datasets focus on frontal, well-illuminated facial images, prioritizing the accurate prediction of a 3D gaze vector in relatively controlled settings. This is in contrast to datasets used in the GTD subfield, which are predominantly 2D and often feature more complex and less constrained scenarios. In these cases, subjects may appear at various distances, and their face may occupy only a small portion of the image, conditions that better reflect real-world use cases but introduce additional challenges for accurate gaze target localization. Another important limitation of current GTD datasets is that they are usually annotated only with 2D information, despite the inherently 3D nature of gaze. Furthermore, most existing approaches rely heavily on facial images as input, often overlooking the privacy concerns that may arise from using identifiable biometric data, an issue that has yet to be fully addressed in the literature.

Our Contribution

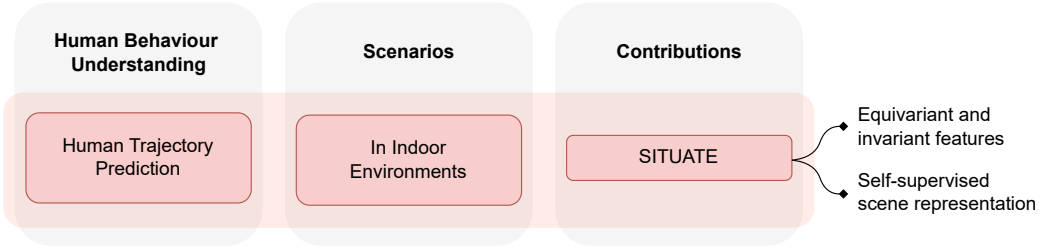
To address the limitations discussed above, particularly those affecting the GTD task, we introduce, in Chapter 5, a novel dataset for 3D visual attention estimation, a new methodology for 3D GTD with privacy preservation as a priority, exploiting the pose of the people, and an application of gaze analysis designed to enhance customer experience in a simulated mall environment. Now that we know how a person will

move and where they will move in the near future, we just have to give an answer to the question: *Why* will they move there?

Gaze is intrinsically linked to a person's activity, often revealing their focus and intent. By analyzing gaze direction and identifying the gaze target within a scene, we gain valuable insights into the person's cognitive state and behaviour, unlocking powerful applications in areas such as HRI, assistive systems, and retail analytics.

Chapter 3

Human Trajectory Prediction *in Indoor Environments*



Patterns of human motion in outdoor and indoor environments are substantially different due to the scope of the environment and the typical intentions of people therein. While outdoor Human Trajectory Prediction (HTP) has received significant attention, indoor HTP is still an underexplored research area.

*This chapter describes **SITUATE**, a novel approach to cope with indoor HTP by leveraging equivariant and invariant geometric features and a self-supervised vision representation. The geometric learning modules model the intrinsic symmetries and human movements inherent in indoor spaces. This concept becomes particularly important because self-loops at various scales and rapid direction changes often characterize indoor trajectories. On the other hand, the vision representation module is used to acquire spatial-semantic information about the environment to predict users' future locations more accurately. We evaluate our method through comprehensive experiments on the two most famous indoor HTP datasets, i.e., THÖR and Supermarket, obtaining state-of-the-art performance. Furthermore, we also*

achieve competitive results in outdoor scenarios, showing that indoor-oriented HTP models generalize better than outdoor-oriented ones.

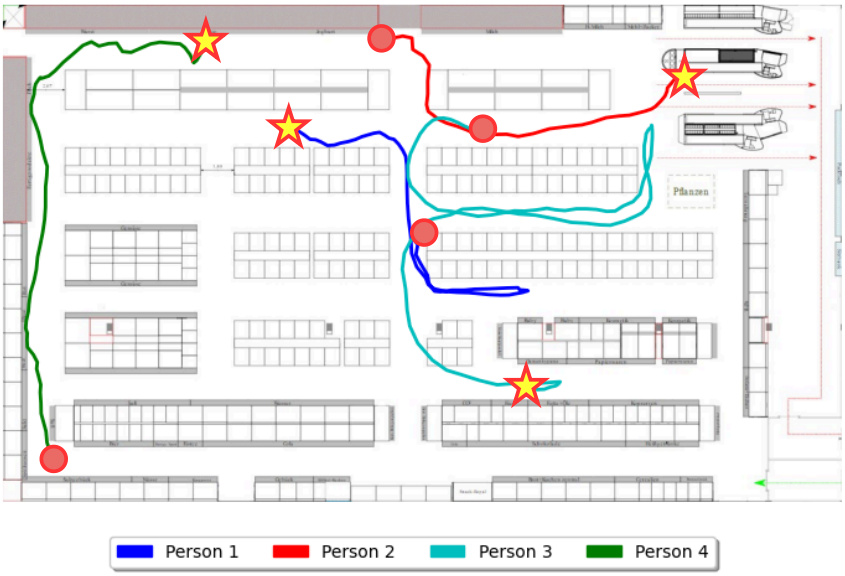


Fig. 3.1 Examples of different trajectories from the Supermarket [66] dataset to show the difficulty of the indoor HTP task. In particular, the dataset showcases long trajectories (*Person 4*), self-loops (*Person 1* and *Person 3*), and confusing movements (*Person 2*) performed in an environment that strongly affects the people's paths. Specifically, the red circle represents the starting point of a trajectory, and the yellow star represents its final point.

3.1 Introduction

Human Trajectory Prediction (HTP) is the task of predicting the likely path that a subject will take to reach its designated endpoint [167]. This predictive process finds its applicability and utility in a multitude of domains [111]. For example, in the context of robotics, it serves as a tool for facilitating the predictions on potential future robot trajectories, useful for intelligent planning considering human responses [168]. In industry, HTP becomes critical for optimizing automated systems and ensuring seamless interactions with other occupants and components of a production line [169].

Despite the significant volume of research over the past decade devoted to outdoor HTP [3, 80, 168, 72, 75, 207, 9], there has been a notable scarcity of studies that exploited user trajectory data in indoor settings [134, 166, 165, 202, 178, 201, 186], also considering the crucial role these predictions play nowadays in the development of location-based services within indoor spaces. This gap in research inspired this work, which investigates a learning framework designed explicitly for indoor HTP.

In Figure 3.1, we can note the distinctive nature of indoor settings, where users can encounter numerous choices and potential pathways. This factor implies that the dynamic of the motion can be strongly influenced by the environment setup [134, 30]. Users can navigate through different interconnected rooms, corridors, doors, and elevators, often having the freedom to deviate from straightforward paths and choose alternative routes. Indoor spaces also have a higher density of structural elements and potential obstacles, such as furniture, walls, and partitions, as shown in Figure 3.1, related to the Supermarket [66] dataset. Outdoor environments provide more open spaces, where visibility is less restricted, and the impact of physical barriers is typically reduced [41].

Consequently, indoor HTP requires a deeper understanding of the context and semantics of the indoor space, as users may have specific goals, like finding a particular room, reaching a specific point of interest, or accessing various facilities [77]. This contextual richness adds a layer of complexity to the prediction process since it also makes it necessary to consider the space's physical layout. Considering the omnipresence of indoor environments in human lives, it is imperative to address HTP in these situations. Indeed, recent studies show that humans spend most of their time in indoor environments such as homes, supermarkets, airports, conference facilities, and train stations [5]. These considerations form the basis of the research conducted in this work.

We present **SITUATE**, the first model designed specifically to cope with indoor HTP by leveraging equivariant and invariant geometric feature learning and a self-supervised vision representation. Taking inspiration from [208], the equivariant and invariant geometric learning modules were employed to accurately represent intrinsic movements, like self-loops at various scales and hierarchies inherent in indoor spaces. On the other hand, the self-supervised vision representation module enabled us to acquire spatial-semantic information about the environment, using the scene or space layout images when available, to predict users' future locations meaningfully and more accurately.

In summary, the main contributions of this research work are:

- We present **SITUATE**, a novel approach for indoor HTP based on equivariant and invariant geometric feature learning modules and a self-supervised vision representation.
- The equivariant and invariant modules are used to cope with the problem related to the more complicated movements inherent in indoor spaces.

- The vision representation module is used to acquire spatial-semantic information about the environment to predict users' future locations more accurately.
- **SITUATE** also achieves competitive results in outdoor scenarios, showing that indoor HTP models generalize better than outdoor-oriented ones.

3.2 Related Work

3.2.1 Indoor Human Trajectory Prediction

In indoor settings, people navigate in loosely constrained but cluttered spaces with multiple goal points that can be reached in many ways [166]. Moreover, people in indoor scenarios tend to focus on their surroundings, fixating on the most interesting parts of the scene, alternating movement and stationary phases [186].

Some works have been proposed to address the specific problem of indoor HTP [134, 166, 165, 178, 201], highlighting the differences between indoor and outdoor trajectories. In [165], the authors address the problem of generalizability, proposing a novel indoor dataset and new metrics to normalize common biases. They tackle the problem of aleatoric multimodality with the GAN-Tri model, which uses a heuristic to produce samples corresponding to different behaviours. [134] examine trajectories, modelled as a Markov Chain, within 3D environments, introducing the concept of an occupancy map to represent the relative accessibility of each point on the map with respect to its geometry. The study emphasizes the importance of proximity from each point to the destination and the occupancy frequency in constructing a probability transition matrix for HTP. Unlike them, our approach considers indoor spaces' detailed scene layouts and non-trivial human movements.

3.2.2 Equivariant and Invariant Graph Neural Networks

Inspired by the research on rotation-equivariant Convolutional Neural Networks within the 2D image domain [48], the advent of Graph Neural Network (GNN) architectures opened doors to investigating symmetries beyond rotations [215]. For example, in [170], the authors proposed partial equivariance by focusing on translation equivariance. Meanwhile, [64] constructed filters using spherical harmonics, enabling equivariance to rotations and translations and facilitating transformations between higher-order representations.

In [171], a new model for learning equivariant GNNs, dubbed EGNNs, is proposed. Differently from the previous works, this formulation maintains the flexibility of GNNs while remaining E(n) equivariant (translation, rotation, and reflection equivariant) without the need to compute expensive higher-order operations. [98] further extended this concept by incorporating geometrical constraints implicitly encoded in the forward kinematics when tackling molecular dynamics prediction and human motion capture. However, a significant limitation of current methods is their focus solely on state prediction, preventing models from effectively using sequence information.

Recently, EqMotion [208] extended on these ideas to propose an equivariant motion prediction parametric network with an invariant interaction reasoning module, able to tackle distinct problems such as particle and molecule dynamics, Human Pose Forecasting (HPF), and outdoor HTP. Interaction invariance is fundamental in ensuring the agents' interactions remain constant under input transformation.

In our research, we adapt some of the concepts presented in [208] to propose an equivariant model for the HTP task, which, in combination with a module to extract semantic information about the scenes, unlock more precise forecasting capabilities in indoor settings.

3.2.3 Self-Supervised Vision Representation

One way to get image representations without heavily relying on annotated data is to perform Self-Supervised Learning (SSL). In a nutshell, SSL learns deep feature representations invariant to sensible transformations of the input data. Then, the learned representations could be used in supervised downstream tasks.

The self-supervised vision representation state-of-the-art rapidly evolved, with Transformer-based architectures emerging as leading models. The Vision Transformer (ViT) [55], and its variants like DeiT [191], have demonstrated impressive performance in learning powerful visual representations from unlabeled data. Specifically, these models leverage self-attention mechanisms to capture global context and long-range dependencies within images, enabling them to encode rich semantic information efficiently.

In this work, to extract semantic information from scenes represented in a 2D map, we use the pretrained BEiT [12], the state-of-the-art self-supervised vision representation model. This offers a powerful framework for learning visual represen-

tations without explicit supervision, effectively capturing high-level semantics and intricate features inherent in visual data.

3.3 Method

Mathematical Background

Given a set of transformations $T_x : X \to X$, a function $F : X \to Y$ is called equivariant if exists a transformation $T_y : Y \to Y$ equivalent to T_x, on the Euclidean space such that:

$$F(T_x(X)) = T_y(F(X)) \ . \tag{3.1}$$

Moreover, we also want the model to have the invariance property. Given the same set of transformations, a function $F : X \to Y$ is called invariant on the Euclidean space if it exists a transformation $T_y : Y \to Y$ such that:

$$F(X) = F(T_x(X)) \ . \tag{3.2}$$

Specifically, this work addresses the problem of multi-person HTP by considering the input trajectories as a graph. As proven by [171], during the message passing of a GNN, the property of equivariance can be ensured by enriching the features of the neighbour nodes with the $L2$ distance between nodes. Let $G = \{V, E\}$ be an input graph representing the input trajectory with nodes $v_i \in V$ and edges $e_{ij} \in E$. For every node v_i, a feature vector $h \in \mathbf{R}^h$ and an absolute position $x_i \in \mathbf{R}^3$ are given. To preserve equivariance among different layers of the model, we update the position as follows:

$$m_{ij} = \phi_e \left(h_i^l, h_j^l, \left\| x_i^l - x_j^l \right\|^2 \right) \ , \tag{3.3}$$

$$x_i^{l+1} = x_i^l + C \sum_{j \neq i} \left(x_i^l - x_j^l \right) \phi_x \left(m_{ij} \right) \ , \tag{3.4}$$

where C is equal to $1/(M-1)$ with M number of nodes, ϕ_e and ϕ_x are learnable Multi-Layer Perceptrons (MLPs), defined as $\phi_e(\cdot) = W_e \cdot + B_e$, l indicates the layer and m_{ij} represents the information passed between two nodes during the message passing.

As reported in [171], ϕ_x has to be a scoring function $\phi_x : X \to S$, with $S \in \mathbf{R}^1$. With this procedure, the update of V is consistent, allowing the model to learn without

being affected by SO(2) transformations, with SO(2) being the group of all rotations in the plane around the origin that preserve the Euclidean norm, mathematically described by 2×2 matrices. Furthermore, the features learned across layers must be consistent and invariant to graph transformations. To do so, the following procedure governs the final message-passing operations and the update of the features carried out by the $i-th$ layer:

$$m_i = \sum_{j \neq i} m_{ij} \,, \tag{3.5}$$

$$h_i^{l+1} = \phi_h \left(h_i^l, m_i \right) \,, \tag{3.6}$$

with ϕ_h, an MLP also designed as $\phi_h(X) = W_h X + B_h$, responsible for the invariant feature learning. Mixing these two components allows us to build an equivariant and invariant GNN using Euclidean SO(2) transformations.

Motion Prediction

Here, we introduce the general problem formulation of motion prediction. We have a multi-agent system with m agents. Each agent is represented as A_i, where $i = 1, 2, \ldots, m$. The goal is to predict the future motions of these agents based on their historical observations. For each agent A_i, we can denote the historical observations as X_i. These observations typically include positions and can be represented as $X_i = \{x_0^i, x_1^i, \ldots, x_t^i\}$, where x_t^i represents the position of agent A_i at time step t. We also add the velocity $S_i = \{s_{t+1}^i, s_{t+2}^i, \ldots, s_{t+f}^i\}$ as input information of the model. The velocity of an agent is a natural invariant feature because it is not affected by any translation or SO(2) transformation. We use the velocity to compute the initial feature vector of a specific agent A_i. More details in Section 3.3.2. Specifically, for each agent A_i we aim to predict its future f positions $Y_i = \{y_{t+1}^i, y_{t+2}^i, \ldots, y_{t+f}^i\}$.

3.3.1 The SITUATE Prediction Network

In this section, we present **SITUATE**, our HTP network that explicitly uses equivariant and invariant geometric features and a self-supervised scene representation module to tackle the indoor HTP problem. The model architecture is shown in Figure 3.2. The first module we present is in charge of producing the scene-representation encoding. As anticipated, the subjects' motion characteristics differ greatly from those of the outdoor case when considering indoor HTP. Indeed, the motion is strongly

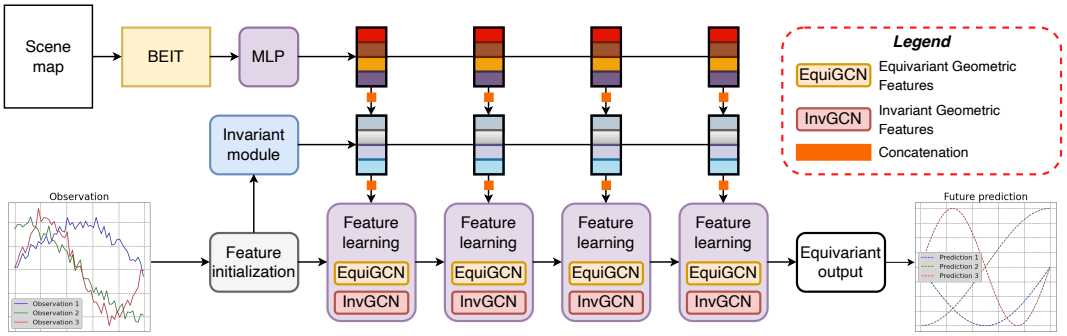


Fig. 3.2 In **SITUATE**, we first produce a feature vector regarding the scene using the self-supervised vision representation module. Then, a feature initialization layer is used to initialize geometric and pattern features. We then successively update the geometric and pattern features by the equivariant geometric feature learning and invariant pattern feature learning layers, obtaining expressive feature representation. We further use an invariant reasoning module to infer an interaction graph used in equivariant geometric feature learning. Finally, we use an equivariant output layer to obtain the final prediction.

characterized and limited by the objects and obstacles in the scene. Knowing the available space that limits the viable paths in the scene can, for every X_i, strongly reduce the cardinality of all the possible outcomes of the model. Starting from the assumption that all the scene objects and structure are available in the form of a scene layout or a camera image, BEiT [12] is first used to output visual tokens T_s, the so-called scene-representation encodings. These tokens T_s are fed into a learnable MLP defined as $\phi_t : T_s \rightarrow T_e$ and then concatenated into the input.

The input concatenated with T_e is then fed into two modules: the Equivariant Graph Convolutional Network (GCN) block (*EquiGCN*) and the Invariant GCN block (*InvGCN*). Following [208], these two blocks are both based on the implementation of the message passing described in Equation (3.3), modified to accept also T_e:

$$m_{ij} = \phi_e \left(h_i^l, h_j^l, \left\| x_i^l - x_j^l \right\|^2, T_e, a_{ij} \right) , \tag{3.7}$$

where a_{ij} is the edge attribute (or weight), which can be derived from the adjacency matrix.

Specifically, the *EquiGCN* block is responsible for the update of the node's coordinates x, and it represents the implementation of the update function described in Equation (3.4). On the other hand, *InvGCN* is the implementation of the update function of the node's features h in Equation (3.6).

The possible pathways are learned by the module ϕ_t, starting from the token produced by the pretrained BeIT model. The outputs of *EquiGCN* and *InvGCN* are computed as reported respectively in Equation (3.4) and Equation (3.6), updating h_i^{l+1} and x_i^{l+1}.

To understand the contribution of these two modules in less formal and more practical terms, imagine a navigation system. It can get from point A to point B but might struggle with tricky situations. In **SITUATE**, the *EquiGCN* module injects a sense of direction like a compass you wear on your hat. Specifically, it ensures the network understands the environment's layout, regardless of where it starts "looking". The *InvGCN* module, on the other hand, acts like a map you hold – it helps the network account for different starting points and body orientations, making the predictions more robust.

3.3.2 Feature Initialization

The input given to our model is a set of trajectories of different agents. The first step is to define a node for every position x_t for every agent A_i. Every node x_t is connected with x_{t-1} and x_{t+1} if they are related to the same agent A_i. Since only trajectories (and thus positions x_t) are given as starting data, it is necessary to define for each trajectory a vector of initial features h_i^0 to be used as input together with the positions x_i^0.

As stated in [98], having an invariant feature vector h_i^0 is necessary to guarantee equivariance. Given that as input data we only have position X_i, we followed the procedure in [208] to use velocities in order to create h_i^0 as follows:

$$\hat{x}_i = \phi_X(X_i + \overline{\mathbb{H}}) + \overline{\mathbb{H}}\,, \tag{3.8}$$

$$\rho_i^t = \|v_i^t\|_2\,, \tag{3.9}$$

$$\theta_i^t = \text{angle}(v_i^t, v_i^{t-1})\,, \tag{3.10}$$

$$h_i^0 = \phi_{h_0}(\rho_i, \theta_i)\,, \tag{3.11}$$

where h_i^0 is the initial features vector of the $i-th$ agent. v_i^t represents the velocity of the agent and is defined as $\triangle\hat{x}_i^t$, where $\triangle$ is the finite difference operator, $\overline{\mathbb{H}}$ is the centroid of the observed trajectories of all agents in the scene. ϕ_X and ϕ_{h_0} are two fully connected layers responsible for encoding and producing the initial graph and the initial features of the trajectory.

To compute h_i^0, two different types of velocities are needed (thus, information invariant to rotation and translation): $\triangle x_i^t$ effectively represents the Euclidean velocity of the agent and θ_i^t represents the angular velocity on a certain time step t. Note that both ϕ_{x_0} and ϕ_{h_0}, and in general all the operations described, are linear transformations: this is necessary to preserve both equivariance and invariance properties of the remaining part of the model.

3.4 Evaluation

Our experimental evaluation is tailored toward two objectives. Firstly, in Section 3.4.2, we show the superiority of **SITUATE** in the two most well-known indoor datasets, defining the new state-of-the-art in indoor scenarios. Secondly, in Section 3.4.3, we prove that **SITUATE** can also achieve comparable results with respect to other competitors on outdoor datasets. Finally, we report some ablation studies in Section 3.4.4.

3.4.1 Evaluation Setup

Datasets

We evaluate **SITUATE** on the state-of-the-art indoor datasets and the most well-known outdoor HTP dataset.

THÖR. The THÖR dataset [166] includes human motion trajectory and gaze data collected in an indoor environment with accurate ground truth for the participants' position. It comprises 395,000 frames at 100 Hz, 2531K people detections, and over 600 individual and group trajectories between multiple resting points. The map was taken from the dataset's official website.

Supermarket. The Supermarket dataset [66] comprises 4 different scenarios: German1, German2, German3, and German4, *i.e.*, four different supermarket. The dataset collection involved attaching devices on shopping carts/baskets and recording their movements during customer usage. Each subset includes a file with a map of the supermarket.

ETH-UCY. The ETH [155] and UCY [117] dataset group consists of five different scenes: ETH & HOTEL (from ETH) and UNIV, ZARA1, & ZARA2 (from UCY). The scenes are captured in unconstrained outdoor environments with few objects blocking the pedestrians' paths. In this case, images of the scenes were used.

Evaluation Metrics

We use standard metrics for the HTP task, *i.e.*, minimum Average Displacement Error (ADE), and minimum Final Displacement Error (FDE). In particular, ADE measures the average $L2$ difference between the prediction at all time steps and the ground truth. On the other hand, FDE measures the difference between the predicted endpoint and the ground truth.

Prediction Mode

Following the evaluation protocol of [208], **SITUATE** is employed in two prediction modes: deterministic and multi-prediction. Deterministic means the model only outputs a single prediction for each input motion observation, while multi-prediction means the model has 20 predictions for each input motion observation. Under multi-prediction, ADE and FDE will be calculated using the best-predicted trajectory. To adapt to multi-prediction, we modify **SITUATE** to repeat the last feature updating layer and the output layer 20 times in parallel to have a multi-head prediction.

Implementation Details

As a backbone for our model, we used the structure of [208]. The model architecture has four layers of geometric feature learning. We use the Sigmoid Linear Unit (SiLU) activation function and dropout with a 0.5 probability to regularise within all MLPs. The visual embeddings of the image, *i.e.*, the floor plans (look at Figure 3.1) for context information are derived from the last layer of the BEiT model. The model is provided with past trajectory information spanning eight discrete time steps, and the model's task is to predict 12 steps into the future. In addition to the dropout mentioned above, we apply the Discrete Cosine Transform (DCT) to the input data as a regularisation technique. Specifically, by representing the data in the frequency domain, it becomes easier to distinguish between signal and noise components, resulting in a cleaner signal. The impact of these regularization approaches is discussed in Section 3.4.4. We train our models with a batch size of 64 for 60 epochs,

Table 3.1 Deterministic HTP performance (ADE(m)/FDE(m)) on the THÖR and the Supermarket datasets. The **bold** and underlined fonts denote the **best** and second-best results.

Deterministic	Parameters	Performance (ADE (m) ↓ / FDE (m) ↓)		
		THÖR	**Supermarket**	**Average**
TransF [72]	799,232	2.62/4.81	2.56/2.90	2.59/3.85
MemoNet [207]	10,654,904	0.78/5.05	1.79/2.94	1.28/3.99
EqMotion [208]	4,340,036	0.56/0.94	1.71/2.65	1.13/1.79
SITUATE	4,351,566	**0.45/0.93**	**1.21/1.84**	**0.83/1.38**

Table 3.2 Multi-prediction performance (ADE(m)/FDE(m)) on the THÖR and the Supermarket datasets. The **bold** and underlined fonts denote the **best** and second-best results.

Multi-prediction	Parameters	Performance (ADE (m) ↓ / FDE (m) ↓)		
		THÖR	**Supermarket**	**Average**
PECNet [133]	1,568,664	–	1.57/3.45	–
GP-Graph [9]	36,287	2.80/3.92	3.19/4.57	2.99/4.24
EqMotion [208]	4,340,036	1.32/1.03	1.29/1.77	2.61/1.40
SITUATE	4,351,566	**0.50/0.86**	**0.53/0.65**	**0.51/0.75**

using AdamW [126] as an optimizer within the PyTorch Lightning framework on an NVIDIA RTX 3090.

3.4.2 Indoor Human Trajectory Prediction Results

We conducted comparative experiments to assess the soundness of our approach against existing HTP methods. The methods include deterministic evaluation models (TransF [72], MemoNet [207]), as well as multi-prediction evaluation models (PECNet [133], GP-Graph [9]). Our evaluation for both prediction modes also encompasses EqMotion [208], the state-of-the-art method with invariant end equivariant interaction reasoning. Table 3.1 and Table 3.2 show the results.

The results show that the proposed **SITUATE** consistently outperforms all baseline methods in all cases. On the THÖR dataset, **SITUATE** achieves an ADE of 0.45 and an FDE of 0.93, showcasing its superiority over other models. Notably, compared to EqMotion, the second-best model, **SITUATE** exhibits a substantial 22% reduction in ADE and a 1% reduction in FDE.

Table 3.3 Deterministic HTP performance (ADE(m)/FDE(m)) on the scenes of the ETH-UCY dataset. The **bold** and underlined fonts denote the **best** and second-best results.

Deterministic	Performance (ADE (m) ↓ / FDE (m) ↓)					
	ETH	**HOTEL**	**UNIV**	**ZARA1**	**ZARA2**	**Average**
S-LSTM [3]	1.09/2.35	0.79/1.76	0.67/1.40	0.47/1.00	0.56/1.17	0.72/1.54
SGAN-ind [80]	1.13/2.21	1.01/2.18	0.60/1.28	0.42/0.91	0.52/1.11	0.74/1.54
Traj++ [168]	1.02/2.00	0.33/0.62	0.53/1.19	0.44/0.99	0.32/0.73	0.53/1.11
TransF [72]	1.03/2.10	0.36/0.71	0.53/1.32	0.44/1.00	0.34/0.76	0.54/1.17
MemoNet [207]	1.00/2.08	0.35/0.67	0.55/1.19	0.46/1.00	0.37/0.82	0.55/1.15
EqMotion [208]	0.96/1.92	**0.30**/0.58	**0.50/1.10**	**0.39/0.86**	**0.30/0.68**	**0.49/1.03**
SITUATE	**0.94/1.90**	**0.30/0.57**	**0.50/1.10**	0.41/0.89	0.32/0.70	**0.49/1.03**

Similarly, on the Supermarket dataset, **SITUATE** continues demonstrating its effectiveness with an ADE of 1.21 and an FDE of 1.84. Compared to EqMotion, again the closest competitor, **SITUATE** achieves a 29% reduction in ADE and a 31% in FDE, reinforcing its dominance.

The consistently good performance of **SITUATE** across both datasets underscores its robustness and efficacy in HTP tasks. The results further suggest that **SITUATE** is accurate and that using scene information when tackling indoor prediction scenarios offers a key advantage compared to the available approaches.

3.4.3 Outdoor Human Trajectory Prediction Results

We also evaluate the performance of **SITUATE** with the deterministic and multi-prediction modalities with outdoor scenarios. Here, we show that **SITUATE** achieves competitive results, showing that indoor-oriented forecasting models tend to generalize better than outdoor-oriented ones. Table 3.3 and Table 3.4 present the quantitative results.

Specifically, when considering the deterministic prediction case, **SITUATE** demonstrates a performance improvement by obtaining state-of-the-art results in both ADE and FDE across the ETH (0.94/1.90), HOTEL(0.30/0.57), and UNIV (0.50/1.10) scenes. It places second in the ZARA1 and ZARA2 scenes while performing on par with EqMotion [208] when considering average performance. In the context of multi-prediction modality, **SITUATE** secures the second rank in terms of ADE and FDE across nearly 75% of the scenes within the ETH-UCY dataset while maintaining an overall second place in average performance.

Table 3.4 Multi-prediction performance (ADE(m)/FDE(m)) on the scenes of the ETH-UCY dataset. The **bold** and underlined fonts denote the **best** and second-best results.

Multi-prediction	Performance (ADE (m) ↓ / FDE (m) ↓)					
	ETH	**HOTEL**	**UNIV**	**ZARA1**	**ZARA2**	**Average**
SGAN [80]	0.87/1.62	0.67/1.37	0.76/0.52	0.35/0.68	0.42/0.84	0.61/1.21
STGAT [99]	0.65/1.12	0.35/0.66	0.34/0.69	0.29/0.60	0.52/1.10	0.43/0.83
STAR [216]	0.36/0.65	0.17/0.36	0.31/0.62	0.29/0.52	0.22/0.46	0.26/0.53
NMMP [95]	0.61/1.08	0.33/0.63	0.52/1.11	0.32/0.66	0.43/0.85	0.41/0.82
Trajectron++ [168]	0.61/1.02	0.19/0.28	0.30/0.54	0.24/0.42	0.18/0.31	0.30/0.51
PECNet [133]	0.54/0.87	0.18/0.24	0.35/0.60	0.22/0.39	0.17/0.30	0.29/0.48
Agentformer [218]	0.45/0.75	0.14/0.22	0.25/0.45	0.18/**0.30**	0.14/0.24	0.23/0.39
GroupNet [206]	0.46/0.73	0.15/0.25	0.26/0.49	0.21/0.39	0.17/0.33	0.25/0.44
MID [75]	**0.39**/0.66	0.13/0.22	**0.22**/0.45	**0.17/0.30**	**0.13**/0.27	**0.21**/0.38
GP-Graph [9]	0.43/0.63	0.18/0.30	0.24/**0.42**	**0.17**/0.31	0.15/0.29	0.23/0.39
EqMotion [208]	0.40/**0.61**	**0.12/0.18**	0.23/0.43	0.18/0.32	**0.13/0.23**	**0.21/0.35**
SITUATE	0.41/0.64	0.13/0.20	0.23/0.43	0.22/0.35	0.14/0.26	0.22/0.37

Table 3.5 Ablation results (ADE(m)/FDE(m)) of **SITUATE**. We assess the contribution of the scene representation module and regularization methods in the deterministic prediction case.

Scene Representation	**Regularization**	Performance (ADE (m) ↓ / FDE (m) ↓)		
		THÖR	**Supermarket**	**Average**
✗	✗	0.50/1.02	1.92/1.55	1.21/1.29
✗	✓	0.56/0.74	1.79/2.94	1.18/1.84
✓	✗	0.57/0.96	1.29/1.89	0.93/1.43
✓	✓	**0.45/0.93**	**1.21/1.84**	**0.83/1.38**

Designed for indoor scenarios and their peculiar conformations, **SITUATE** remarkably demonstrates robust capabilities, even when tested on outdoor datasets. In contrast, this is not always true for architectures tailored for outdoor instances, as we can observe in Table 3.1 and Table 3.2, that often struggle when confronted with scenes that differ from those for which they were designed.

3.4.4 Ablation Study

We quantitatively evaluate the impact of the scene representation module and regularization methods by considering the deterministic indoor prediction scenario. Results are summarized in Table 3.5. It is observed that both contributions play a

crucial role in enhancing the overall performance of **SITUATE**. In particular, the scene representation module effectively encodes semantic information from the visual scene maps, facilitating an accurate understanding of the environment. On the other hand, the regularization methods ensure robustness and generalization of the model by mitigating overfitting and improving its ability to generalize to unseen data. Therefore, we assert that both contributions are indispensable for achieving the desired outcomes and validating the efficacy of our approach.

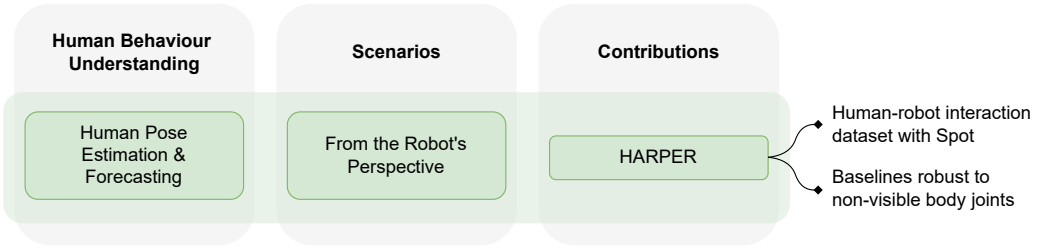
3.5 Findings

This chapter presents **SITUATE**, a GNN-based model designed specifically to cope with indoor HTP. **SITUATE**, using geometric features and self-supervised vision representations, models the intricate human movements inherent in indoor spaces and accurately predicts users' future locations. The scene vision representation module provides insights about the environment, particularly helping in those indoor scenes that are more constrained and full of obstacles. We evaluate our method on two well-known indoor HTP datasets, *i.e.*, THÖR and Supermarket, and achieve state-of-the-art prediction performance. Furthermore, we also achieve competitive results in outdoor scenarios, showing that indoor-oriented forecasting models generalize better than outdoor-oriented ones.

Future works on this topic should focus on testing different environments with various levels of complexity, from the arrangement of the obstacles to the difference w.r.t. the people's flow inside them. This could lead to the development of more robust methodologies, that can adapt to a whole range of different use cases, even those that integrate both indoor and outdoor layouts. Another interesting research direction is represented by the study of the motivations that guide human movement. Understanding whether a person is moving toward a specific target or engaged in a particular activity can provide a critical context for interpreting predicted trajectories. Finally, the long-term goal is to implement this type of prediction algorithm inside a fully vision-based pipeline, where the starting point is just the video coming from a camera, without prior annotations of human positions.

Chapter 4

Human Pose Estimation and Forecasting *from the Robot's Perspective*



This chapter introduces ***HARPER****, a novel dataset for 3D Human Pose Estimation (HPE) and 3D Human Pose Forecasting (HPF) in dyadic interactions between users and Spot, the quadruped robot manufactured by Boston Dynamics. The key novelty of* ***HARPER*** *is its focus on the robot's perspective, i.e., on the data captured by the robot's sensors. This makes 3D body pose analysis challenging, as being close to the ground results in only partial captures of humans, differently from what happens in standard applications of 3D HPE. The scenario underlying* ***HARPER*** *includes 15 actions, of which 10 involve physical contact between the robot and users. The corpus contains recordings not only from Spot's built-in stereo cameras but also from a 6-camera OptiTrack MoCap system. The recordings from both acquisition systems have been synchronized. This setup leads to ground truth skeletal representations with a precision of less than a millimeter. Additionally, the*

dataset includes reproducible benchmarks for 3D HPE, 3D HPF, and Collision Prediction (CP), all based on publicly available baseline approaches. In particular, being able to estimate only the visible joints of the human pose, those not visible are reconstructed using a diffusion-based model. This enables future ***HARPER*** *users to rigorously compare their results with those provided in this work.*

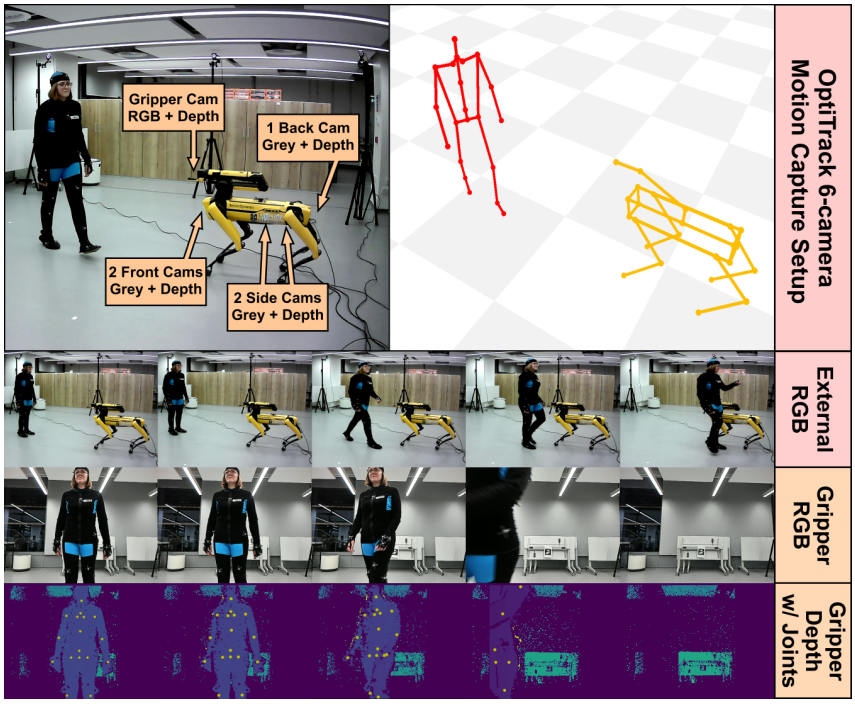


Fig. 4.1 HARPER showcase. (Top-left) We exploit the Spot on-board equipment to let the robot perceive people. (Top-right) Thanks to a 6-camera OptiTrack MoCap setup, we capture 3D human poses represented with 21 joints and 0.035 mm of error. (Second row) An additional external RGB camera shows the actions performed. (Third row) The gripper RGB camera PoV. (Fourth row) The gripper depth camera PoV, with the ground truth joints in yellow.

4.1 Introduction

One of the main changes characterizing the transition from Industry 4.0 to Industry 5.0 is the shift from Human-Robot Interaction (HRI) to Human-Robot *Collaboration* [130]. This shift requires robots to evolve into cobots, *i.e.* intelligent platforms equipped with capabilities like visual perception, action recognition, intent prediction, and safe online motion planning. These technologies empower cobots with awareness of their surroundings, enabling them to adapt their behaviour in real time, a stark contrast to the rigid, pre-programmed routines of traditional robots [57]. In other words, understanding human behaviour is a crucial requirement for a robot to become a cobot and, correspondingly, to be capable of adaptive and seamless interaction with its users [200].

Thus motivated, we propose Human from an Articulated Robot PERspective (**HARPER**), a new, publicly available dataset revolving around the interaction

between human users and Spot, the quadruped robot manufactured by Boston Dynamics. Such a platform attracts increasingly more attention, and, not surprisingly, it was recently included in Habitat 3.0 [157], one of the most advanced environments for simulating HRIs. Moreover, Spot is an ideal cobot candidate for at least three reasons: 1) the four-legged design and the biologically inspired locomotion enable it to operate on diverse and challenging terrains (the robot can even climb stairs [23]), thus making Spot a potential companion in a wide range of settings [143, 76, 84, 145]; 2) Spot is equipped with one of the most advanced self-balancing systems available on the market, significantly limits the risk of accidents during close physical interaction with users; and 3) Spot is equipped with a total of 5 greyscale + depth sensors mounted on its body and an RGB-D camera on its grasper arm (see Figure 4.1). This is important because such a sensing apparatus makes Spot particularly suitable for analysis and understanding of human behaviour, a critical step in the evolution from robot to cobot.

HARPER includes dyadic interactions between Spot and 17 human users, 5 females and 12 males, each performing 15 actions that require different degrees of collaboration with the robot (see Section 4.3 for more details). The data captured with the Spot sensors were enriched with the recordings of a 6-camera OptiTrack Motion Capture (MoCap) system capable of extracting skeletal models of the users. The joints were localized with a precision of less than one millimeter (see Figure 4.1), thus providing highly accurate ground truth information about the pose and position of the users. This is a significant advantage because Spot's sensors and MoCap cameras are synchronized. Therefore, skeletal models can be used to reliably validate approaches for human behaviour analysis and understanding based solely on Spot sensors.

In addition to the above, skeletal representations enable one of the key novelties of **HARPER**, namely the possibility to train approaches capable of recognizing 3D body pose and movement when the Spot, due to its limited height, can "see" its users only partially, something that happens whenever the distance between them is small. To the best of our knowledge, this is one of the first datasets that allows the investigation of such a problem in 3D.

We asked the 17 **HARPER** participants to stage two types of physical contact with the robot, namely *unintentional* and *intentional*, according to the terminology proposed in [63]. The first type includes (staged) collisions, while the second includes punches, kicks and soft touches. We paid special attention to the first type because of the major role collisions play in scenarios based on co-located interac-

tions. We enriched **HARPER** with benchmarks, *i.e.*, reproducible experimental protocols and baseline approaches designed to address three tasks relevant to the analysis of physical contact, namely 3D Human Pose Estimation (HPE) (especially when Spot can "see" its users only partially), 3D Human Pose Forecasting (HPF) and Collision Prediction (CP). This allows researchers interested in **HARPER** to rigorously compare their results with those presented in this article (see Section 4.4).

Overall, the main contributions of this work can be summarized as follows:

- We propose the first dataset that includes not only the Point of View (PoV) of the robot (the data captured with the Spot's sensors) but also a panoptic PoV (the data captured with the MoCap system) that provides accurate ground truth information for position and pose of both users and robot.
- To the best of our knowledge, **HARPER** is the first dataset enabling the reconstruction of the human users' pose with the data captured with a quadruped robot, a problem which is challenging because Spot is short and the subject is usually close to it (hence, the cameras cannot capture the whole body of the user).
- **HARPER** allows, for the first time, the visual prediction of collisions between a mobile robotic platform and users.

The rest of this chapter is organized as follows: Section 4.2 surveys previous work, Section 4.3 describes **HARPER** in detail, Section 4.4 presents the benchmarks, and Section 4.5 draws the conclusions.

4.2 Related Work

Table 4.1 shows the main differences between **HARPER** and existing datasets of similar scope. Most available corpora are based on the analysis of people's trajectories. The THÖR dataset [166], a well-known example, contains the 2D trajectories of 9 human users moving together with a robot. Besides this, the data includes 6D head positions, LiDAR data from a stationary sensor, and the participants' orientations and eye gaze directions. THÖR-Magni [172], a significantly more extensive dataset from the authors of THÖR, introduces on-board sensors on the mobile robot and semantic attributes describing the roles and activities of detected people. Similarly, the JRDB [139] dataset aims to enable mobile robots to detect

Table 4.1 Main HRI datasets revolving around human movement and its analysis. Values in the *Participants* column indicated with the asterisk (*) refer to datasets captured in uncontrolled scenarios.

Dataset	Participants	Actions	Mobile Robot	Robot PoV	Human Skeleton	Human Joints	Marker-Based MoCap	Robot Skeleton	Collisions / Intended Contacts
THÖR [166]	9	13	✓	✗	✗	✗	✓	✗	✗
THÖR-Magni [172]	40	48	✓	✓	✗	✗	✓	✗	✗
JRDB [139]	3.5K*	N/A	✓	✓	✗	✗	✗	✗	✗
L-CAS Multisensor [213]	N/A*	N/A	✓	✓	✗	✗	✗	✗	✗
FROG [4]	1M*	N/A	✓	✓	✗	✗	✗	✗	✗
CODa [221]	N/A*	N/A	✓	✓	✗	✗	✗	✗	✗
PTUA [225]	N/A	N/A	✓	✓	✗	✗	✓	✗	✗
InHARD [49]	16	14	✗	✗	3D	17	✗	✗	✗
JRDB-Pose [196]	5K*	N/A	✓	✓	2D	17	✗	✗	✗
HuRoN [90]	N/A* (5/17 for exp)	N/A	✓	✓	✗	✗	✗	✗	✓
NatSGD [176]	18	11	✗	✓	estim. 2D	25	✗	Arm	✗
CHICO [169]	20	7	✗	✗	2D, 3D	15	✗	Arm	✓
SCAND [105]	N/A* (14 for exp)	12	✓	✓	✗	✗	✗	Quadruped, Wheeled	✗
UF-Retail-HRI [36]	8	2	✓	✓	3D	23	✗	Arm	✗
HARPER	17	15	✓	✓	2D, 3D	21	✓	Quadruped	✓

and track humans in both indoor and outdoor settings. The data includes stereo cylindrical RGB videos and LiDAR point clouds annotated with 2D and 3D bounding boxes. In addition, the dataset includes benchmarks for both 2D and 3D detection and tracking. A more recent version includes 2D human-pose skeletal annotations [196].

Other datasets provide information about objects that the robots can encounter while moving. For example, CODa [221] aims at both object detection and semantic segmentation. It was acquired with a wheeled robot, and it features sequences in indoor and outdoor settings of a university campus, as well as 3D semantic segmentation and 3D object detection benchmarks. In the case of FROG [4], based on LiDAR sensors placed on a wheeled robot at roughly the height of human knees, the problem is the detection of people in possibly crowded sites where humans can be confused with static and dynamic obstacles. A similar issue is at the core of the dataset proposed in [225], where the material is collected with an RGB-D camera mounted on a small mobile robot. The annotations include attributes such as, *e.g.*, the presence of static obstacles, illumination and human poses. An OptiTrack MoCap system provides information about the position of both the robot and users. The problem of navigating through an environment, possibly shared with humans, is the focus of HuRon [90]. The data was collected using a Roomba bot equipped with LiDAR, bumper collision detectors, video, and odometry sensors. However, pose annotations about the people sharing the space with the robot are not provided.

HARPER shows major novelties with respect to the datasets above. The availability of 3D skeletons for both the robot and users provides unprecedentedly detailed information about the interaction between the two, especially when considering that the joints are captured with a precision of less than 1 millimeter thanks to the Opti-

Track MoCap system. A similar acquisition precision is achieved with InHARD [49], an industrial HRI dataset featuring both RGB images and MoCap data of a person performing multiple manual tasks, captured with wearable devices. A robotic arm, mostly stationary, is the platform used for the experiments, and it never collides with the user, offering a looser type of interaction. This is not the case in **HARPER**, which includes physical contacts of different types. In [36], a mobile wheeled robot is employed to capture an HRI dataset in a retail environment. Multiple people navigate the room and perform picking and sorting actions while the robot moves along with them. Egocentric and scene videos, eye gaze directions, point clouds, and other data are collected. The human poses are collected through an IMU-based MoCap device, which requires careful setup and calibration for every person. However, the Spot used in **HARPER** is a more advanced robotic platform, and its movement is significantly less constrained.

Skeleton representations were also used in other datasets. In [169], the scenario is a collaboration between a user and a robotic arm in an industrial setting. A MoCap system captures the user's skeleton from an external PoV, missing the robot's perspective. Furthermore, the acquisition is markerless, making the joint localisation less precise. In another dataset, the multimodal NatSGD [176], the goal is imitation learning, and the data includes human commands, such as speech and gestures, with a focus on robot behaviour in the form of synchronized robot trajectories. However, the joint localization is, once again, less precise than in **HARPER** because it is performed by applying the HPE algorithm Openpose [29] to the videos.

Finally, to the best of our knowledge, the only other dataset in which the robot Spot was actually involved is SCAND [105], where two robots, a wheeled one and the Spot, are teleoperated in human-populated environments. A large variety of data is acquired thanks to an additional LiDAR sensor mounted on the two robots. However, no skeletal models are considered for humans, a major difference with respect to **HARPER**. The dataset we propose appears to have distinctive characteristics with respect to those currently available in the literature.

4.3 The HARPER Dataset

The primary motivation behind the design of **HARPER** is to expand the research opportunities enabled by previous HRI datasets (see Section 4.2), especially consid-

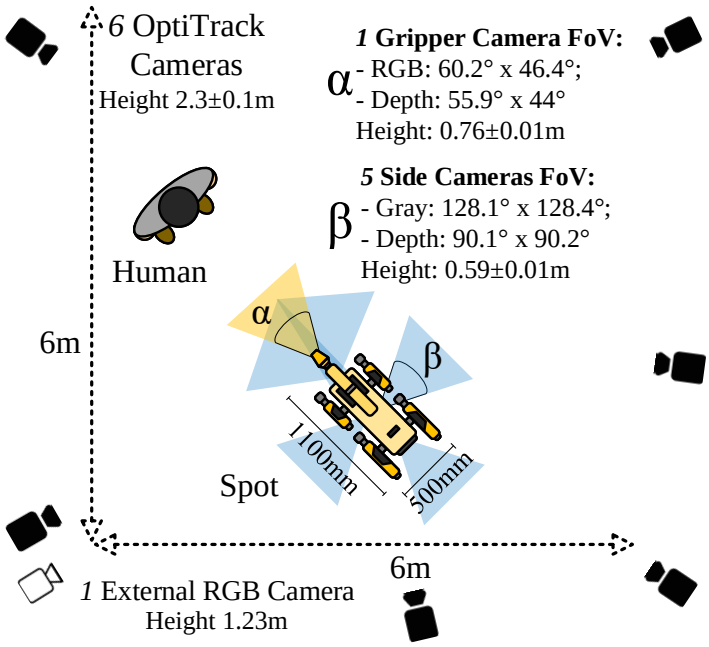


Fig. 4.2 A 6-camera OptiTrack MoCap system covers a 6×6 squared meters area where users and Spot can freely move. The external RGB camera's FoV covers the setting. The 5 Spot on-body greyscale + depth cameras and the RGB-D frontal camera (gripper) cover the environment surrounding the robot.

ering the transition from robots to cobots. The collection of the corpus involved 17 participants who were asked to perform 15 actions (the same for all participants). The data was captured with the sensors equipped on Spot: 5 greyscale + depth sensors and one RGB-D camera mounted on the gripper. Moreover, we used 6 MoCap sensors (OptiTrack) and one RGB camera capturing the full setting (see Figure 4.2).

Overall, **HARPER** contains 607 sequences for a total of over 60,000 RGB images, greyscale images, depth frames, and 3D data from multi-sensor recordings. In the following, we discuss the acquisition setup (Section 4.3.1), we describe the actions we captured and their annotations (Section 4.3.2), and, finally, we provide key statistics about the data (Section 4.3.3).

4.3.1 Acquisition Setup

We collected all data in a laboratory (the layout is in Figure 4.2). The 6 cameras of the OptiTrack MoCap system were arranged to cover a 6×6 m^2 area, free of obstacles, in which the participants performed the 15 actions of the **HARPER** scenario. All participants wore a MoCap suit with 37 reflective markers distributed according to the

OptiTrack *Baseline Marker Set* configuration. After calibration, the OptiTrack tracks the markers with a 0.035 mm error at a sampling frequency of 120Hz. Furthermore, thanks to the configuration above, the OptiTrack software (Motive) automatically extracts a 21-joint skeleton representation based on the marker positions.

The robot involved in the experiment is the Boston Dynamics Spot, a 12-Degrees of Freedom (DoF) (3 per leg) quadruped robot equipped with 5 stereo cameras (greyscale with depth) around its body and one RGB-D camera on the gripper. The Spot acts within the OptiTrack area described above, and its skeleton is obtained by applying forward kinematics to its internal motors state, acquired through the API provided by Boston Dynamics. The Spot skeleton is positioned in the same 3D scene as the participants' skeletons using a 4-marker rigid body mounted on its back and tracked by the OptiTrack.

The Spot cameras operate at about 10 Frames Per Second (FPS), and their data is synchronised with the OptiTrack data. This ensures one of the most distinctive features of **HARPER**, *i.e.*, the availability of two viewpoints, one from the robot and a panoptic view that covers the whole scene. The synchronization is achieved by aligning the timestamps of the data with a temporal alignment error of less than 2 milliseconds. It is worth noticing that the overlap between Spot cameras is limited to the 3 frontal cameras with a partial overlap. As a reference, we added an external RGB camera, positioned outside the OptiTrack delimited area, to capture the whole scene (see bottom left part of Figure 4.2). All the videos recorded with such a camera are provided with the dataset.

4.3.2 Actions and Annotations

We invited 15 university students to participate in the data collection (5 females and 12 males). They all signed an informed consent letter, and the information they provided, including the data collected during their participation, was handled according to the ethical regulations of the university in which the data was collected. Each participant interacted with the Spot individually in a session that followed a consistent sequence of steps. First, participants were assisted in wearing the suit necessary for marker tracking, and then they were asked to display a T-pose for calibrating the skeleton extraction.

After calibrating the OptiTrack, we asked the participants to perform 15 actions designed to reproduce different situations (Table 4.2), with the robot standing still for 8 actions and moving for the other 7 actions. In particular, the participants were

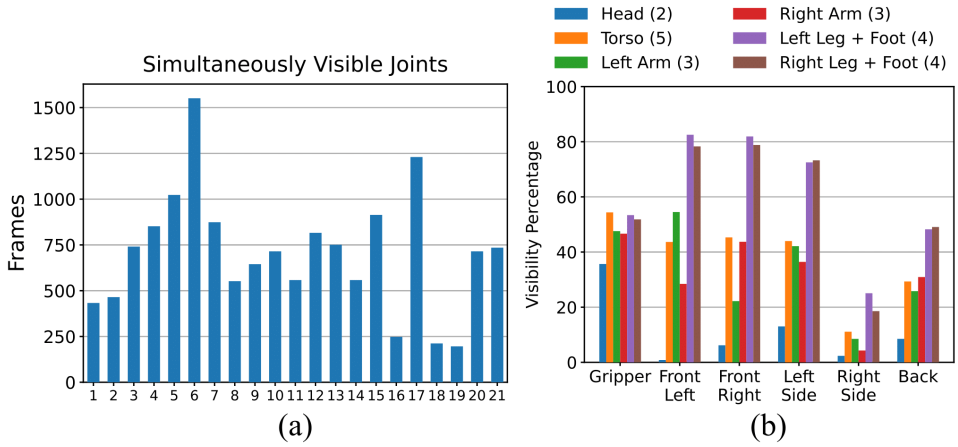


Fig. 4.3 Joints visibility from the robot's perspective. The left chart shows how many frames contain exactly n joints for $n = 1, \ldots, 21$. The right plot shows the percentage of frames in which the different parts of the skeleton are visible.

instructed to act collisions as realistically as possible, *i.e.*, as if they were accidentally and unintentionally bumping into the Spot. The area covered by the OptiTrack is sufficiently broad to perform the actions comfortably. However, some participants moved inadvertently out of it, leading to missed markers in a few frames. Similarly, some occlusions prevented the OptiTrack from working properly for a few moments. These issues concerned no more than 3% of the total frames, and missing markers were effectively replaced through linear interpolation, ensuring that the skeleton representation was acquired with continuity and with the same precision at all times.

OptiTrack and Spot share the same reference system, which enables the projection of 3D skeletons onto the videos captured with the robot's cameras (greyscale and RGB). This allows for accurate annotation of the joint positions in the video. In addition, given that the robot's leg motor state is known, forward kinematics was applied to compute the position of the robot's joints in the 3D space, starting from the rigid body mounted on the Spot back. This allowed us to obtain a 21-joint representation of both the participants' skeletons and the robot's skeleton.

4.3.3 Dataset Statistics

Figure 4.3(a) shows, for all possible values of n, the number of frames in which exactly n human skeleton joints are visible to the robot. This information is important for understanding the difficulty level involved in addressing one of the new tasks that **HARPER** enables, namely the analysis and understanding of human poses when

Table 4.2 HARPER actions. The expression *Contact* means that the distance between Spot and user is lower than 10 cm.

Action	Action Description	Robot Moving	Contact
A1 Walk+Crash Frontal	Human walks towards Spot oriented frontally then collides;		✓
A2 Walk+Crash 45°	Human walks towards Spot oriented at 45° then collides;		✓
A3 Walk+Crash Sideway	Human walks towards Spot oriented at 90° then collides;		✓
A4 Walk+Crash Backwards	Human walks towards Spot oriented backwards then collides;		✓
A5 Walk+Stop	Human walks towards Spot, then stops right before colliding;	✓	
A6 & A7 Walk+Avoid	Human and Spot walk towards each other avoiding collision at last second on the right (A6) / left (A7).	✓	
A8 Walk+Touch	Human walks towards Spot, then physically touch it;		✓
A9 Walk+Kick	Human walks towards Spot, then kicks it;		✓
A10 & A11 Walk+Punch	Human walks towards Spot oriented at 0° (A10) / 90° (A11), then punches it		✓
A12 Circular Walk	Human and Spot walk together in a circular path	✓	
A13 Circular Follow +Touch	Human follows Spot in a circle, then touches it with the hand	✓	✓
A14 Circular Follow + Avoid	Spot follows the human in a circle, then avoids contact	✓	
A15 Circular Follow + Crash	Spot follows the human in a circle, then a collision happen	✓	✓

they are only partially visible. Similar information is shown in Figure 4.3(b), where human joints are grouped according to five body parts, *i.e.*, head (2 joints), torso (5 joints), left/right arm (3 joints), and left/right leg (4 joints). The figure reports

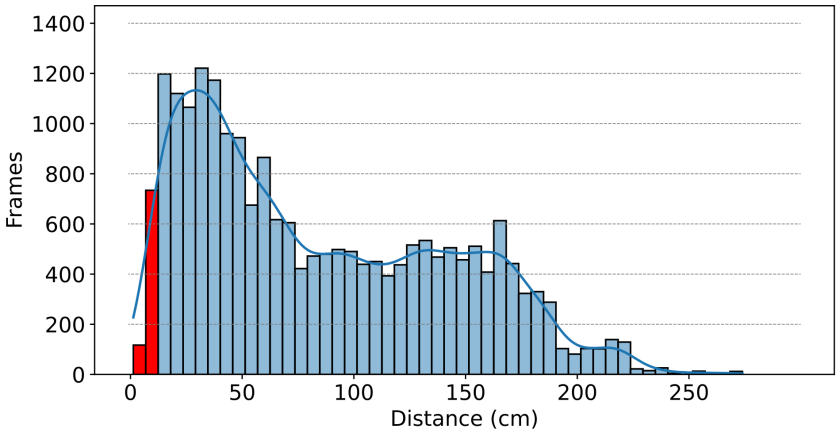


Fig. 4.4 Distribution of distances between Spot and users (the distance considers the two closest joints of human and robot). Red columns correspond to distances lower than 10 cm, considered as cases of physical contact.

the percentage of times these body parts are visible (one joint is sufficient for the part to be considered visible) to each camera. One of the main patterns is that the gripper camera is more likely to capture the upper part of the body and legs, but not the feet (*i.e.*, the spike on 17 visible joints caused by the gripper camera Field of View (FoV)).

Regarding the interaction between Spot and the participants, Figure 4.4 shows the histograms of the distances between their closest joints. Two modes appear: below and above a distance of 1.3 meters. Distances corresponding to physical contact are highlighted in red. A threshold distance of 10 cm was used to determine whether physical contact was happening (see details in Section 4.4.3).

4.4 Evaluation

HARPER provides three benchmarks: *3D HPE*, *3D HPF*, and *CP*. All benchmarks are from the *robot's perspective*, *i.e.*, they are based on data captured with the robot's sensors, which is one of the key novelties of **HARPER**. Participants S1-S12 were used for training (15,984 frames), while participants S13-S17 were used for testing (5,542 frames). For 3D HPF, we sampled 10,990 sequences (of 20 frames each) for the training set and 3,502 for the test set, keeping the same distribution of participants. The sequences were sampled by using a rolling window with a 1-frame step. Sequences without visible joints were excluded from the test set.

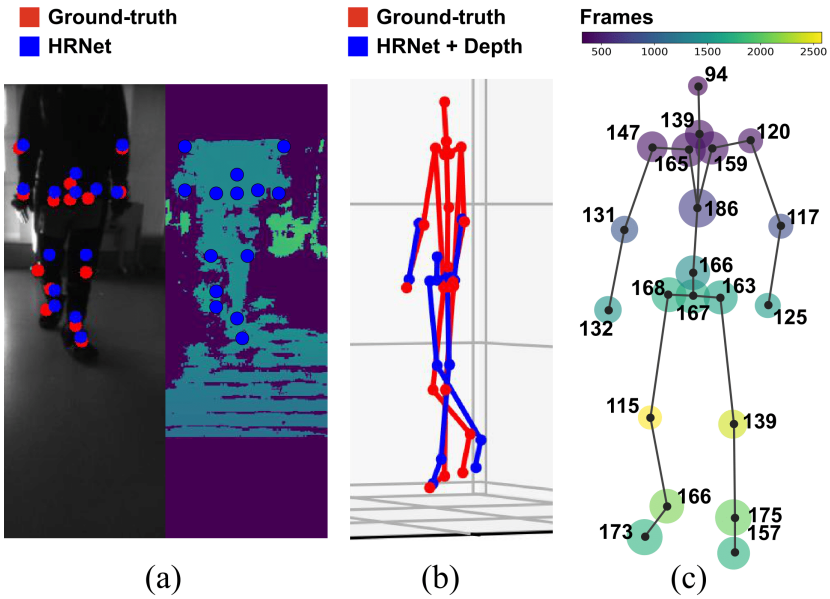


Fig. 4.5 Results of 3D HPE from the robot's perspective. (a) On the left, the predicted 2D joints (in blue) by HRNet [181] and the corresponding ground truth joints (in red). On the right is the depth image with the same 2D detections. The depth will serve to do the lifting. (b) The lifted 3D poses alongside the complete OptiTrack skeletons. (c) MPJPE (in mm) for every visible joint (inside the depth FoV) on the test set. The size of the blobs is proportional to the errors, while colours are related to the number of times a joint is visible from the robot's perspective.

4.4.1 3D Human Pose Estimation

3D HPE from the robot's perspective involves finding the 3D coordinates of the visible human joints using greyscale images and synchronized depth maps captured by the robot's sensors. The main challenge is that humans may not be fully visible (Section 4.3). The proposed baseline approach first uses a 2D pose estimator to find the position of the visible joints and then computes their 3D positions by exploiting the depth values, as shown in [123] (Figure 4.5).

The 2D pose estimator is HRNet [181], trained on **HARPER** training data after resizing the images to 256×256 (no augmentation is applied). The depth sensors' FoV is narrower than that of the video cameras. Therefore, if an estimated joint is out of the depth FoV, it is considered non-visible. The positions of the joints, along with their corresponding depth values, can then be mapped into the 3D OptiTrack coordinate system. Once this task is completed, the 3D points inferred by the approach can be compared with those of the MoCap ground truth skeleton.

Table 4.3 3D HPF errors. We provide the MPJPE expressed in mm with a prediction horizon of 400 and 1000 ms. The errors are computed for the particular frame for each action (first nine columns) as well as the average over all frames (*Average*) and the average over the last frame of each action instance (*Last Frame Average*). We use three different inputs: the ground truth OptiTrack skeletons (GT), the incomplete poses estimated by the 3D HPE baseline described in Section 4.4.1 or the filtered ground truth poses where only the joints seen by Spot were kept. *Depth* (D) means the poses are computed using the acquired depth maps. *Reconstruction* (R) means a trained diffusion-based imputation model reconstructs the poses before passing them to the forecasting method.

Actions		**A1-4**		**A5**		**A6-7**		**A8**		**A9**		**A10-11**		**A12**		**A13**		**A14**		**A15**		**Average**		**Last Frame Average**	
Time (ms)		400	1000	400	1000	400	1000	400	1000	400	1000	400	1000	400	1000	400	1000	400	1000	400	1000	400	1000	400	1000
STS-GCN [2]	GT	115	169	104	150	99	134	116	129	143	218	144	213	148	229	179	268	112	148	150	224	120	171	147	260
	GT+R	182	229	121	153	281	340	126	145	148	213	154	224	156	234	193	288	262	302	306	338	190	242	212	326
	HRNet+D+R	414	469	188	249	559	674	276	306	197	283	257	340	204	294	337	430	646	684	560	606	**382**	453	**400**	**538**
SiMLPe [3]	GT	63	150	60	140	42	99	31	72	66	145	77	183	89	208	99	221	46	108	90	216	60	141	98	264
	GT+R	158	231	100	174	261	351	87	121	82	163	114	199	112	226	137	260	231	302	280	353	158	237	193	359
	HRNet+D+R	436	528	213	275	674	917	290	368	232	349	310	451	205	307	412	610	681	820	785	1170	441	595	511	790
EqMotion [4]	GT	42	110	37	91	25	61	23	60	40	104	60	139	66	174	73	164	35	96	68	168	**41**	**104**	**69**	**197**
	GT+R	147	200	81	125	255	326	79	108	61	124	102	169	92	188	112	185	221	286	267	328	**146**	**205**	**171**	**299**
	HRNet+D+R	419	488	194	234	569	660	274	299	187	237	262	319	187	285	340	407	632	678	613	628	386	**446**	422	552

We evaluated 2D HPE performance with the Percentage of Correct Keypoints (PCK) [214], *i.e.*, the fraction of correct predictions within a distance threshold τ (set to 0.5 on the predicted heatmaps). For the 3D joints estimation, we used the Mean Per Joint Position Error (MPJPE) [103] metric, *i.e.*, the mean Euclidean distance between the visible estimated joints and the ground truth OptiTrack ones. We obtained a PCK of 86.8% and an average MPJPE (computed only on the visible joints) of 151 mm on 2D and 3D poses, respectively (Figure 4.5). The 2D baseline performs well, especially when considering that, in many cases, only one limb is visible or the participant is very close to the Spot. The 3D lifting shows some limitations due to the noise in the depth maps, especially when the participants are far from the Spot. However, the performance was sufficient to address 3D HPF and CP, both from the robot's perspective. We think that the result we obtained may pave the way for the application of other methods that make better use of the intrinsic characteristics of the various sensors of the Spot, from neural networks that directly exploit depth to methods that use only the video stream and the accurate 3D annotation provided by OptiTrack.

4.4.2 3D Human Pose Forecasting

3D HPF from the robot's perspective involves predicting the future poses of the human subject using the robot's sensors. The pose at time t can be denoted as

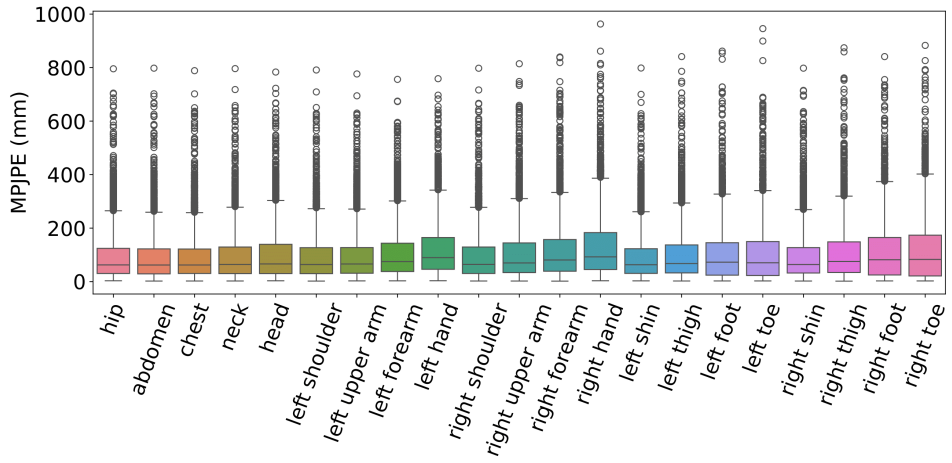


Fig. 4.6 MPJPE for each joint using EqMotion with GT as input and a forecasting horizon of 1000 ms.

$X_t \in \mathbb{R}^{D \times J_h}$, where $D = 3$ is the dimension of the space and $J_h = 21$ is the total number of joints in the human skeleton (X_t is the set of all joint positions in 3D). Correspondingly, 3D HPF means predicting $X_{t+1:t+K}$ based on $X_{t-T+1:t}$, where $X_{i:j} = X_i, X_{i+1}, \ldots, X_j$, and K is the *horizon*. In line with widely-used experimental protocols [100, 180], we set $T = 10$ and $K = 4$ (roughly 400 ms) or $K = 10$ (roughly 1000 ms), two cases referred to as *short-term* and *long-term* forecasting, respectively. We used average MPJPE over the K predicted frames (average MPJPE) or MPJPE over the K^{th} predicted frame (final MPJPE) as performance metrics. The pose forecasting baselines we applied are STS-GCN [180], SiMLPe [78] and EqMotion [208]. All three were trained using MPJPE as a loss function without applying augmentation. The training was performed using the 21-joint poses obtained with the OptiTrack sensor as a ground truth.

Each baseline has three variants based on different assumptions about the input data: *GT* assumes that the robot can access all ground truth joints in the human skeleton, *GT+R* assumes that the robot can access only the joints visible to its sensors, and *HRNet+D+R* represents the 3D pose as described in Section 4.4.1. The GT+R and HRNet+D+R baselines deal with input sequences of incomplete poses. These cannot be processed with the forecasting baselines above and, in general, with any of the approaches in the literature. Therefore, we used a diffusion-based time series imputation model, CSDI [184], built on a cascade of transformer blocks with skip connections. This model takes a sequence of incomplete poses as input and uses

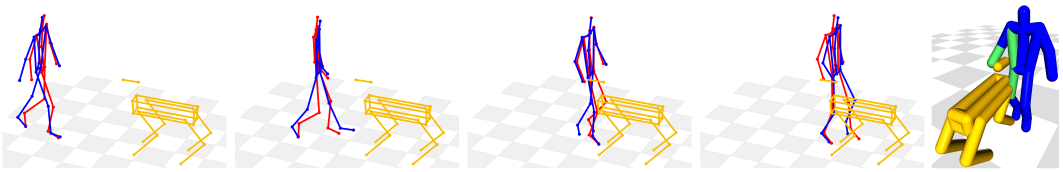


Fig. 4.7 Qualitative results for the 3D HPF with the 1000 ms horizon. (a) shows the human pose forecasted in blue and the ground truth in red. At the end of the sequence, an accidental collision occurs. In (b), the collision (highlighted in green) is detected as explained in Section 4.4.3. The forecasting approach used is EqMotion [208] on the GT data.

them to condition the generation of a complete pose, reconstructing the position of missing joints.

Table 4.3 shows the results for the three variants of every baseline. GT poses perform best, while HRNet+D+R, corresponding to the most challenging task, performs the worst. EqMotion [208] delivers the best absolute results when in the presence of GT data: 43 mm, on average, over the 400 ms horizon and 70 mm over the 1000 ms horizon. However, STS-GCN [180] bridges the performance gap with EqMotion when the data is noisier like, *e.g.*, in the HRNet+D+R case: the best average MPJPE is 313 mm over the 400 ms horizon and 332 mm over the 1000 ms horizon, while EqMotion achieves an MPJPE of 309 mm over the 400 ms horizon, and of 333 mm over the 1000 ms horizon.

Finally, we computed the MPJPE for each joint using the baseline with the smallest average error, *i.e.*, EqMotion [208], with GT as input (Figure 4.6). We also estimated the correlation r between these errors and the average velocity of each human joint with the Pearson coefficient ($r = 0.79$, $p = 1.79e-05$), noticing that the faster a joint moves, the harder it is to predict its trajectory in the future.

4.4.3 Collision Prediction

CP from the robot's perspective is the task of predicting whether the user and the robot will have physical contact, regardless of intent. We focus on the contact or collisions caused by humans with the robot due to the intricate challenges associated with predicting human movements, especially when the body is partially visible. Table 4.2 shows that **HARPER** includes four types of physical contact, all acted out to the best of the participants' abilities. Since they differ significantly in terms of energy and limbs involved, we addressed them as different cases in the experiments. The CP process takes as input a sequence of human poses $X_{t+1:t+K}$ and a sequence

Table 4.4 Performance of the different Collision Prediction methods with a 1000 ms horizon in terms of *Accuracy*, *Sensitivity*, and *Specificity* scores. The evaluation is divided into the four categories of contacts included in the **HARPER** dataset.

Method	Input Type	Unintended			Touch			Kick			Punch		
		Acc.↑	Sen.↑	Spec.↑	Acc.↑	Sen.↑	Spec.↑	Acc.↑	Sen.↑	Spec.↑	Acc.↑	Sen.↑	Spec.↑
STS-GCN [180]	GT	0.93	0.92	**0.94**	0.92	0.91	0.93	0.85	0.78	0.90	0.78	0.83	0.71
SiMLPe [78]	GT	0.93	0.93	0.92	0.94	0.92	**0.97**	0.81	0.83	0.79	0.78	**0.89**	0.63
EqMotion [208]	GT	**0.95**	**0.96**	**0.94**	**0.95**	**0.96**	0.94	**0.93**	**0.88**	**0.96**	**0.82**	0.83	**0.82**
Depth-based	D	0.49	0.25	**0.90**	0.53	0.33	0.87	0.71	0.39	**0.94**	0.60	0.61	0.59
EqMotion [208]	HRNet+D+R	**0.74**	**0.61**	**0.90**	**0.89**	**0.88**	**0.91**	**0.80**	**0.73**	0.86	**0.67**	**0.66**	**0.69**

of robot poses $Y_{t+1:t+K} = (Y_{t+1}, \ldots, Y_{t+K})$, where $Y_t \in \mathbb{R}^{D \times J_r}$ ($D = 3$ and J_r is the number of joints of the robot). The sequence $Y_{t+1:t+K}$ is assumed to be known because the robot plans its actions in advance. The goal is to determine whether $X_{t+1:t+K}$ and $Y_{t+1:t+K}$ indicate a physical contact, defined as two cylinders of radius $r = 5.0$ cm centered around the skeletal links of Spot and user are closer than a threshold $d_h = 10.0$ cm (Figure 4.7(b)). Performance metrics used are Accuracy, Sensitivity, and Specificity [89]. Sensitivity is $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$ (measuring how effectively the system avoids False Positives), while Specificity is $\mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$ (measuring how effectively the system predicts True Negatives).

We started the CP experiments by feeding the methods of Section 4.4.2 with the OptiTrack ground truth data. This gave us an upper bound of the performance and showed that punches and kicks are the most difficult to predict (Table 4.2) due to the speed and energy involved. In contrast, the algorithms tested on the touch collision type, with the lowest speed and energy, showed the best performance. Then, we replaced the ground truth data with the pose forecasts computed by EqMotion [208] in its HRNet+D+R variant, previously reconstructed by the CSDI [184] diffusion process (Section 4.4.2). Table 4.2 shows a performance decrease, but it is not significant.

Finally, we evaluated a straightforward baseline called *Depth-Based* in Table 4.4, showing that CP requires more sophisticated approaches. The baseline is a linear regression over the future K depth frames given T previous frames. This allowed us to test whether any points are predicted to get closer than d_h. We set $T = 10$, $K = 10$, and $d_h = 10$ cm, as used in the pose forecasting baselines. As expected, the Depth-Based method performed worse than the other methods, except for kicks, where Accuracy and Specificity were higher, possibly because the robot's cameras capture users' legs more easily than other body parts.

4.5 Findings

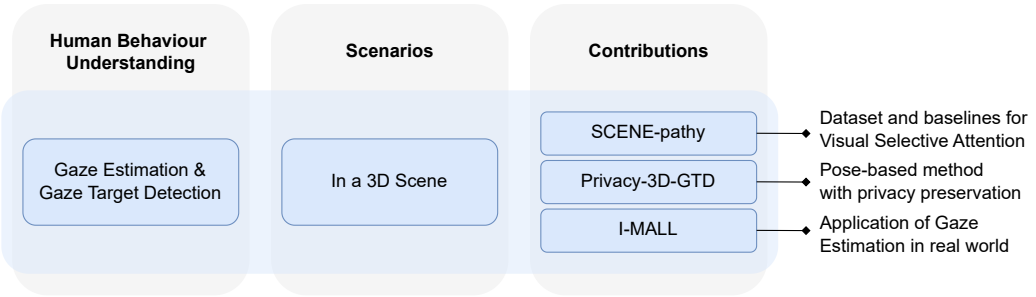
In this chapter, we presented **HARPER**, the first dataset focused on how quadruped robots "see" and predict the movements of human users close to them. The data includes video and depth streams captured with the sensors of Spot, and ground truth skeleton representations of users and Spot performing different interactions, captured with an OptiTrack MoCap system. The interaction scenarios were designed around specific problems and tasks, such as 3D HPE and 3D HPF (Section 4.4). However, the data enables one to address a much wider spectrum of problems, including, *e.g.*, proxemic behaviour estimation [146] and action recognition [45].

In all cases, the key novelty w.r.t. the other datasets in the literature is that the Spot sensors can capture only part of the user's body. This leaves open the challenging problem of reconstructing the full 3D skeleton of the users while having at disposition only a partial 2D image of them. To the best of our knowledge, this is the first corpus revolving around such a problem, and therefore, we enriched the data with benchmarks, including reproducible protocols and baseline approaches. In this way, the experiments we presented can be replicated and the results of future works can be rigorously compared with those obtained in our research.

Possible future works involve using additional state-of-the-art techniques to address all three benchmarks. While our goal for this work was to propose a dataset and a first complete pipeline for CP, it is clear that all three benchmarks can also be considered independently, offering a new way to test how robust the methods in the literature are w.r.t. a different viewpoint. In particular, for the HPE benchmark, it might be interesting to understand the difference between a 2D HPE method plus the depth maps acquired from the Spot sensors and a full 3D HPE methodology. Additionally, future iterations of the dataset should aim to capture more complex human actions, such as multi-person interactions or extended real-world tasks, and include data from diverse quadruped robots with varying sensor configurations and physical characteristics, such as the camera height from the ground.

Chapter 5

Gaze Estimation and Gaze Target Detection *in the 3D Space*



This chapter presents three solutions related to gaze analysis in 3D environments, addressing some of the needs highlighted in the introduction of this thesis: the lack of datasets for more specific case studies and the need to make the algorithms more privacy-friendly. First, we present ***SCENE-pathy****, a dataset and a set of baselines to study people's Visual Selective Attention (VSA) towards the 3D scene in which they are located. In practice, VSA allows discovering which parts of the scene are most attractive for an individual. Capturing VSA is of primary importance in the fields of marketing, retail management, surveillance, and many others. So far, VSA analysis focused on very simple scenarios: a shop shelf or a tiny room, usually with a single subject involved. Our dataset, instead, considers a multi-person and much more complex 3D scenario, specifically a high-tech fair showroom presenting machines of an Industry 4.0 production line, where 25 subjects took part in the experiments, observing the scene and having social interactions.*

Next, we present a novel approach to tackle the problem of 3D Gaze Target Detection (GTD) by utilizing the person's upper-body pose and available depth maps to extract a 3D gaze direction and employing a multi-stage or an end-to-end pipeline to predict the gazed target. When predicted accurately, the human body pose can provide valuable information about the head pose, which is a good approximation of the gaze direction, as well as the position of the arms and hands, which are linked to the activity the person is performing and the objects they are likely focusing on. We demonstrate state-of-the-art results on the most comprehensive publicly accessible 3D GTD dataset without requiring images of the person's face, thus promoting privacy preservation in various application contexts.

*Finally, we briefly present a real-time and real-world industrial solution, where a Gaze Estimation algorithm has been integrated within the **I-MALL** ICT project to improve the customer experience in a shopping hall by tracking the users' interests and profiling their activity to offer a personalised visit.*

5.1 Introduction

Human communication relies on a complex interplay of verbal and non-verbal cues. Among non-verbal cues, the gaze holds particular significance as it reveals where a person directs their visual attention, offering insights into their interests, intentions, or imminent actions [58]. The ability to interpret gaze is crucial for understanding the reciprocal focus and facilitating interactions like turn-taking in social contexts [21].

Gaze analysis spans a broad spectrum of disciplines, including, among others, Human-Robot Interaction (HRI), Neuroscience, Social Robotics, and Organizational Psychology [151, 50, 1, 32]. An analytical approach to estimating the gaze direction or the gazed-at location not only enhances our understanding of human behaviour but also informs the development of technologies that rely on human-machine interaction, social understanding in robotics, and even psychological research into interpersonal dynamics and attentional processes [159, 56, 20]. Thus, the study of gaze represents a multidisciplinary endeavour that continues to evolve with technological and cognitive science advancements.

In computer vision, the study of gaze behaviour primarily involves two main tasks: *Gaze Estimation* and *Gaze Target Detection (GTD)*. Gaze Estimation typically implicates predicting where a person is looking in a three-dimensional space. This means determining the direction of a person's gaze in terms of yaw, pitch, and sometimes roll angles, which correspond to horizontal, vertical, and depth dimensions, respectively. Gaze Estimation can be performed from a single image or a sequence of images, often focusing on the person's eyes or the entire head region [40, 79, 107]. GTD [160] (also referred to as gaze-following), on the other hand, focuses on identifying the specific point or region in the scene that the person is looking at. This is typically done in a two-dimensional space (2D), where the goal is to localize the gaze target's exact coordinates (x, y) within the image or video frame.

Traditional methods for Gaze Estimation historically relied on geometric eye models and regression functions to establish a mapping from eye or face images to gaze vectors [127, 148, 194]. These methods demonstrated effectiveness in controlled environments with consistent subject characteristics, head positions, and lighting conditions; however, their performance diminishes in more diverse and less controlled settings [227]. In recent years, some methods incorporated broader facial features such as head pose and overall facial appearance to enhance accuracy and indicated that leveraging full-face information tends to yield more precise gaze estimates compared to methods focusing solely on eye images [229, 112]. To the

best of our knowledge, no research has explored the use of full-body or upper-body pose rather than head orientation or head/eye images for 3D Gaze Estimation or 3D GTD in unconstrained scenes.

When addressing GTD, the predominant approach has focused on performing this task within a 2D context, utilizing a dual-pathway deep architecture. This architecture typically incorporates inputs such as head crops and scene images [43, 121, 160, 161]. In this setup, one pathway is dedicated to learning feature embeddings from scene images, while the other pathway models head crops corresponding to the individual whose gaze direction is being predicted. Some studies have expanded upon this framework by introducing temporal data integration [44] and incorporating a third pathway to handle depth maps of the scene [59, 102, 144, 188]. Recent advancements, such as [13], have enhanced this pipeline by simultaneously integrating depth maps, 2D human poses, and point clouds. Further innovations incorporated transformers [193, 189] and Vision-Language Models (VLMs) [82]. While numerous attempts have been made to perform GTD in 2D, significantly fewer studies have focused on 3D GTD. However, it has been demonstrated that gaze-following in a 3D scene with only a single camera is feasible [203, 26, 97].

In this chapter, we present three solutions addressing Visual Selective Attention (VSA) estimation, Gaze Estimation, and GTD in the 3D space. In Section 5.3, we present a dataset for estimating the VSA of people toward specific elements in a scene, *i.e.*, the process of directing the gaze to relevant visual stimuli while ignoring the irrelevant ones [33]. VSA can be treated as a special and more complicated case of GTD, where the goal is not only to find the point looked at in the scene at a certain instant of time (*e.g.*, a video frame) but to track how the attention of one or more people has distributed over time on a well-known 3D scene. VSA is generally studied by exploiting the eye gaze dynamics [25, 119] in very simple 2D or 3D scenarios, where the area of interest is limited to planar scenes, such as shop shelf [17, 198], or inside tiny rooms [175, 96]. Applications in contexts such as museums [16] and healthcare [52] could certainly benefit from understanding what people are observing during an extended period (a few minutes) of time. The proposed dataset, dubbed **SCENE-pathy**, is enriched with baselines that exploit Social Signal Processing (SSP) insights to better map the attention onto a 3D scene model.

In Section 5.4, we go back to the more traditional version of GTD, in which the goal is to find a specific point of the environment looked at by a person. The

advantage of the presented methodology is that it can find the point in both the 2D frame and 3D by considering the point cloud of the scene. In addition, what makes our method preferable to other state-of-the-art solutions is the use of the upper-body skeleton to estimate gaze direction rather than face appearance, adding a layer of privacy preservation to the entire pipeline.

Finally, in Section 5.5, we present a real-time and real-world application of Gaze Estimation and GTD, where a video-based methodology for gaze analysis has been implemented as part of a smart shopping mall experience, alongside other framework services such as re-identification, clothing analysis and emotion recognition, to offer a tailored and innovative customer experience.

5.2 Related Work

We provide an overview of the standard GTD task, usually defined in the 2D space, in Section 5.2.1, and the recent developments when it comes to tackling it in the 3D space in Section 5.2.2. Section 5.2.3 gives a brief overview of how attention is mapped onto the considered scene, while Section 5.2.4 expands on the use of SSP methodologies in the study of human attention.

5.2.1 Gaze Target Detection in 2D

GTD is a complex vision task that consists of finding what an observed human subject is looking at, usually by analyzing only an image or a video of the scene from a third-person perspective. [160] formalized it, proposing a 2D dataset named GazeFollow, with multiple target annotations for each image and a double pathway model that extracts features from a subject's head appearance and the scene's salient areas.

This structure has been adopted by multiple following works [43, 121, 160, 161, 44]. For instance, [44] proposed a spatiotemporal modelling of the GTD problem and a way to classify whether the gaze target was inside or outside the scene image. Among these, some introduced a third pathway to incorporate the depth map of the scene image, which is derived using a monocular depth estimator [59, 102, 144, 188]. To overcome the out-of-frame issue, another paper [120] proposed an architecture operating on 360° images, effectively proposing a new challenge. While the approaches outlined in [160, 44, 59, 102, 144, 188] have proven effective, they

are unable to simultaneously detect gaze targets for all the individuals within a scene. This limitation is addressed by Transformer-based gaze target detectors, as shown in [193, 189]. On the flip side, a notable drawback of [193, 189] is their requirement for relatively larger training datasets, causing them to underperform in a low-data regime.

The work of [81] presented the possibility of using skeleton data and depth maps to solve this task, but their method underperformed compared to earlier work. Note that their implementation uses skeleton data represented as images, whereas our methodologies utilize normalized key points (*i.e.*, not images).

5.2.2 Gaze Target Detection in 3D

Although various efforts have been made to extend 2D GTD to the 3D space [13, 93] by estimating depth maps and three-dimensional gaze cones, these methods have been tested only on datasets [160, 44] that provide 2D data and annotations. Similarly, the estimation of VSA [186] involves mapping attention directly onto the 3D point cloud of a scene by analyzing video sequences of individuals moving within a constrained environment. One of our approaches, discussed in Section 5.3, uses 2D inputs but performs its evaluation entirely in 3D space, focusing on how multiple subjects' visual attention maps onto the scene over time. [203] is one of the first methods to deal with this task in the 3D scenario, addressing the attention prediction challenge and investigating why the subject is looking there and what action is being performed.

In [96], an RGB-D dataset with 3D annotations has been proposed, offering manually annotated 3D gaze targets and adding a novel 3D pathway to the now standard saliency-head prediction pipeline. Recently, [97] presented an architecture with four modules for estimating gaze directions, perceiving the stereo Field of View (FoV), generating a 2D gaze heatmap, and retrieving the 3D gaze target. This method utilizes a pretrained ResNet50 as the backbone for the Gaze Estimation module, which takes a cropped head image as input and outputs a 3D gaze unit vector. From the 2D prediction, the 3D target is then retrieved in the evaluation phase. A dataset called GFIE is also presented in the paper, representing the most comprehensive data collection to date for GTD in 3D.

5.2.3 Visual Attention Mapping

One of the distinctive features of the method presented in Section 5.3 is the use of the environment point cloud as a heatmap, with a colouring scheme that shows the areas of major interest. In [16], the visual attention analysis in a museum is carried out with a simple subject-artwork association without considering the environment as a whole. The works of [163] and [17] exploit an external Point of View (PoV) of the scene and propose the concept of a coloured attention heatmap applied to the environment. Differently from us, the used maps are not fine-grained, the frameworks did not consider the interplay of motor activities nor social interactions and discriminate only a fixed discrete number of head orientations, while we use continuous orientations.

5.2.4 The Social Signal Processing Point of View

Humans sample the visual world by making eye movements to direct a centralized region of visual acuity towards different parts of the environment [142]. This allows to direct awareness towards relevant stimuli while ignoring irrelevant ones in the environment and is called VSA [73]. Eye gazing is the clearest way to capture the VSA [62]. Without eye-tracking capabilities due to low-resolution images, the head pose is the second best indicator of where eyes may be fixated [153]. This happens due to the orbital reserve [65], the tendency for the eyes to be positioned centrally in the orbits, and the fact that people orient their heads towards important features in the environment [62]. VSA and vigilance have important interplay during activities such as walking [61] and interacting with other people [197]: it is known that without walking or the presence of social interaction, eye gazing is most probably associated with VSA [73].

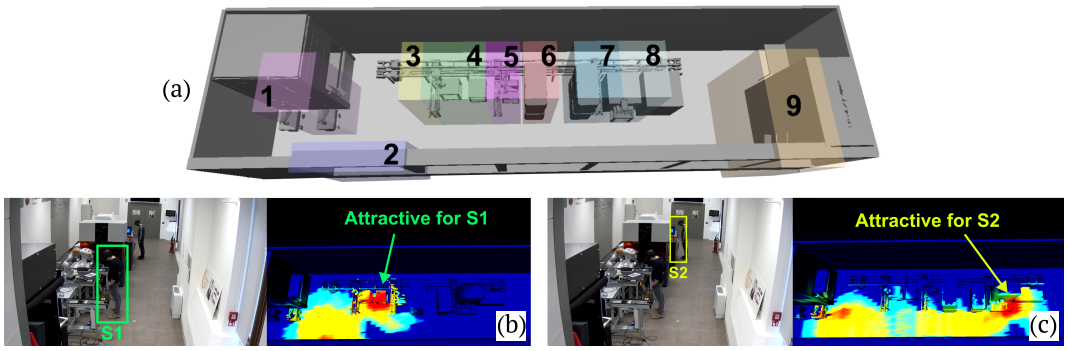


Fig. 5.1 (a) The 3D model of the **SCENE-pathy** showroom scenario, with the nine possible areas of interest highlighted by coloured bounding boxes. (b, c) On the left is an input video frame in which we highlight a single subject. On the right, the corresponding VSA estimation is in the form of a 3D map, where hotter colours indicate areas that attracted more VSA.

5.3 Visual Selective Attention Towards the 3D Scene

In this work, we present **SCENE-pathy** [186], a dataset collected in a 20 by 5 meters showroom of a tech fair (Figure 5.1), where nine different work areas are distinguishable, each one equipped with costly hardware to be exhibited. The showroom has been manually reconstructed beforehand as a 50K-points 3D point cloud. A total of 25 unacquainted subjects have agreed to participate in building the dataset. For each participant, we have collected video sequences and sensory data, thanks to the use of Hololens 2 devices, for about 1.5 hours of data footage. Hololens 2 allowed us to extract tracklets of 3D head positions and eye gaze trajectories, which we collected as annotations for the dataset.

As an additional, higher-level annotation, the subjects filled out a simple questionnaire, communicating which parts of the scenes attracted their attention the most. A dataset with these characteristics would be a novel and useful asset to study problems related to VSA estimation and GTD in general, both in the 2D and 3D spaces. To demonstrate some applications on the dataset, we propose a set of baselines to capture the VSA of single and pairs of people. In general, these baselines individuate and track individuals by 3D Human Pose Estimation (HPE), using solely the RGB video stream as input. Successively, they intersect a physiologically plausible view frustum, fitted on the estimated head pose, with the 3D point cloud (+ 3D culling) of the scene. The resulting intersections are considered as a vote of VSA. Across time, votes are accumulated, providing a 3D weighted map that, for each person,

indicates which parts of the scene have attracted their visual selective attention the most (see Figure 5.1(b), Figure 5.1(c)). The use of a coloured 3D point cloud heatmap is a novel concept in this type of research. The baselines are enriched by SSP findings [17, 197], merging computer vision techniques with social sciences. It is widely known that while moving, the visual attention is aligned preferably with the motion vector [86]. Furthermore, social interactions require visual attention, since humans are naturally interested in observing the other interactants to capture their body language [22]. Put simply, to deal with these aspects in a principled way, our baselines weigh less VSA votes when the person is moving or engaging in social interactions.

In summary, our contributions are the following:

- We introduce **SCENE-pathy**, a dataset with more than 100K RGB frames with associated, for each person, the 3D head positions and the 3D gaze vectors captured by Hololens 2 devices.
- We propose a set of baselines capturing VSA from RGB frames in a complex multi-person scenario, taking into account also the SSP point of view.
- We describe a new way of encoding visual attention using a coloured 3D point cloud, manually reconstructed to allow the VSA counting even on areas occluded to the camera.

5.3.1 The SCENE-pathy Dataset

SCENE-pathy is the first dataset for VSA estimation that features accurate 3D annotations of the position and gaze direction of multiple people moving in the scene, acquired automatically by head-mounted Hololens 2 devices. The main goal is to provide a novel benchmark to study whether VSA can be captured in a non-collaborative scenario using computer vision techniques on a monocular video stream, considering also the SSP point of view.

To this end, we captured two kinds of ground truth data: *low-level* and *high-level*. We define low-level data as the annotations for the subjects' 3D head position and gaze vector relative to a fixed starting position and captured using Hololens 2 devices (one for each person). These ground truth data were then synchronized with the RGB video frames, captured from a single wall-mounted camera, and the starting position was shifted according to the environment coordinate reference system. Since

the main challenge we want to explore in this work lies in capturing perceptual-cognitive processes, *i.e.*, how SSP findings can help reduce false positives in the VSA estimation, we also introduce high-level ground truth in the form of simple questionnaires, in which the subjects have to rank, in descending order of perceived interest, all the nine stations in the scene. The scene is a $20l \times 5d \times 4h$ meters showroom with nine different industrial stations, replicating a modern Industry 4.0 production line (see Figure 5.1(a)).

We manually created an accurate 3D model of the environment using Blender, sampled it into an equally distributed point cloud, and defined 3D bounding boxes around each machine. This will be useful to measure each object's final VSA. The biggest station measures, in meters, $2.5l \times 3d \times 3.5h$, and the smallest measures $0.3l \times 0.1d \times 0.6h$. The dataset is organized into two experimental sessions: *unsupervised* and *supervised*, depending on whether the subjects freely explored the environment or were directed to a specific industrial station. Note that the supervised experiments were performed after the unsupervised ones. Each session was further divided into single-person and multi-person. The multi-person experiments were conducted with two people simultaneously, in which we required them to engage in a discussion, hence having a social interaction.

Unsupervised experiments: In the single-person unsupervised experiments, the subjects were asked to freely explore the environment and then complete the questionnaire. In the multi-person variant, the couple was also expected to discuss their station of interest, motivating their choices, before completing the questionnaire. The average running time of each experiment is about 2 minutes.

Supervised experiments: In the supervised experiments, for each subject, we selected beforehand a specific industrial station to inspect. Obviously, the subjects are now more familiar with the environment, as they have already seen it in the unsupervised session. In this case, we want to be sure which station should be the most looked at to see if we can estimate it correctly. As in the unsupervised session, for the multi-person variant, we also asked the subjects to have a brief discussion after the exploration. Since the subjects can't undergo the unsupervised trials more than once, this second type of test is useful to further enrich the dataset and provide more straightforward and less randomized movement patterns.

The **SCENE-pathy** dataset contains 19 unsupervised experiments (13 single-person and 6 with pairs) and 60 supervised experiments (48 single-person and 12 with pairs). For the data collection, we hired 25 unacquainted subjects: 23 males

and 2 females, with an average age of 27. The dataset comprises the 3D model of the scene and more than 100K RGB frames with associated, for each person, the accurate annotation of 3D head positions, the 3D gaze vectors, and the questionnaires as additional annotation of the actual interest of the people.

5.3.2 The Proposed Baselines

This section presents a modular baseline dubbed *Socio-Dynamic VSA* (*SD-VSA*) to compute the VSA on the **SCENE-pathy** dataset. Furthermore, we provide simpler ablative versions to demonstrate the effectiveness of the involved SSP modules. The first module of the pipeline estimates the 3D pose of each subject in the form of a 3D skeletal model. The captured videos are processed by a top-down 2D HPE algorithm [181], with a tracking module [231] to maintain the subject IDs. The extracted 2D pose joints are then converted by a 3D pose lifter [154] to 17-joint 3D skeletal models, centered initially at the axis origin. These skeletons are then located in their correct scene locations using a homography matrix to transform points from the camera space to a top-view planimetry of the environment. The vector passing through the two 3D joints on the head, one for the nose and the other between the ears, becomes the central axis of a pyramid-shaped view frustum. This represents the direction and area of focus of the subject's gaze. As shown in Section 5.3.3, this estimated vector aligns sufficiently well with the ground truth sensor provided by the Hololens 2 devices in terms of both pan and tilt angular errors.

At the start, each of the N 3D points of the point cloud representing the environment, dubbed P, is assigned a null attention value. For each frame t, and for each human subject i, the head position and the derived gaze direction are used to estimate the subset of visible points F_t^i. This is done in two steps: first, the occluded points are determined and removed using the Hidden Point Removal operator [106]. Then, a view frustum culling is performed considering a horizontal FoV of 90° and a vertical FoV of 60°, mainly based on [113]. In the last step, the attention values of the non-occluded points inside the frustum are increased according to a weighting scheme, with a maximum around the central axis and exponentially decreasing at the margins, following the idea of the attention spotlight model [156]. The view frustum is quantized into three concentric parts with different weights to speed up the computation.

Specifically, the VSA_t^i for a specific frame t and a subject i is composed by the indexes of the point cloud with an associated attention value $s_{t,k}^i = \gamma$, if $p_k \in F_t^i$, else

0, with p_k the k-th point of P. The weight γ is quadratic and is chosen depending on which subdivision of the view frustum contains the considered point. The weights can then be further modulated according to the subject's movement and social activity, as described in the following subsections. Considering the interval of frames T and the subject i, the individual map VSA^i is the sum of the map scores at each frame t: $\text{VSA}^i = \sum_t^T \text{VSA}_t^i$.

The Dynamic Module

Section 5.2.4 explains that motor activity and social interaction override a subject's visual attention towards the environment [61, 62]. In the *Dynamic* module of *SD-VSA*, the motor activity is modelled by changing the equation for $s_{t,k}^i$, substituting γ with $(\gamma \cdot w(v))^2$, where $w(v)$ is a weight inversely proportional to the velocity of the subject in the scene. In simple words, the higher the velocity, the lower the weight that a specific intersection between the view frustum and the scene does have.

The Social Interaction Module

This module copes with the fact that social interactions inhibit the VSA of the individuals towards the scene elements [22, 197]. Many approaches are available in the literature; the most effective ones consider proxemics cues (the position of the interactants) and kinesics cues (body poses). In particular, [17] individuates a social interaction when people are close enough (accounting to Kendon's findings on Hall's usage of the personal space [27]). In practice, people at a distance of less than 1.2 meters who are pointing their faces at each other indicate a social interaction.

All these cues are available: thanks to the multi-person HPE, this configuration can be found simply by analyzing each possible pair of stationary individuals. On each pair, if the people are close enough, two conditions are checked: if there is an intersection between the view frustums and if each one of the vision cones contains the other subject's head. When in the presence of both conditions, the counting of the scene-frustum intersections is suspended.

Ablative Baselines

Alternative baselines can be obtained by suppressing either one or both of the social signal modules. The *Vanilla* baseline indicates the scenario in which none of the modules is active, and the VSA maps are always increased by the same amount, no

matter the speed of the subjects or the occurrence of social interactions. *Dynamic* and *Social* represent, respectively, the baseline in which only the *Dynamic* module is active or only the *Social* module is active. Of course, the *Social* module only makes sense in experiments with more than one subject.

5.3.3 Evaluation

This section presents the experiments and comments on the results obtained by applying our baselines. Firstly, we show that our 3D head pose estimation pipeline (approximating the gaze) performs sufficiently well, demonstrating quantitatively the orbital reserve effect [65] discussed in Section 5.2. Since all of the baselines of Section 5.3.2 depend on the gaze direction, this is a necessary analysis step. Then, we show the capability of *SD-VSA* and the alternative baselines in providing accurate VSA maps on the **SCENE-pathy** dataset, despite the uncertainty in the true gaze direction. Finally, we show that our baseline can provide useful VSA maps also in more crowded scenarios, using the GVEII benchmark [11].

Capturing the Human Gaze

We extracted the Hololens 2 head poses and compared them with the eye-tracking data by measuring the pan and tilt angles error, resulting in less than 1° and 3° discrepancy, respectively. Then, to assess the quality of our head pose estimation pipeline, we calculated the pan and tilt error of the whole dataset in the range [0°, 360°] for the pan and [0°, 180°] for tilt. For pan, 70% of the elements are in the range of 0° and 180°, while on tilt, 97% of the elements are between 60° and 135°. Globally, the average pan error is 52.36° and the average tilt error is 50.81°. Given the size and complexity of the analyzed scenario, we consider these results acceptable.

Unsupervised Experiments

In this section, we show how VSA maps associated with each subject are coherent w.r.t. the high-level ground truth questionnaires discussed in 5.3.1, and the impact of the *Dynamic* and *Social* modules detailed in Section 5.3.2. We start by comparing the ranking from the questionnaires with those obtained by the computed VSA scores. In particular, the nine objects are ranked, taking the maximum score within the bounding boxes defined around each station. Subsequently, we compute the Cumulative Matching Curve (CMC) over all the experiments, comparing the two

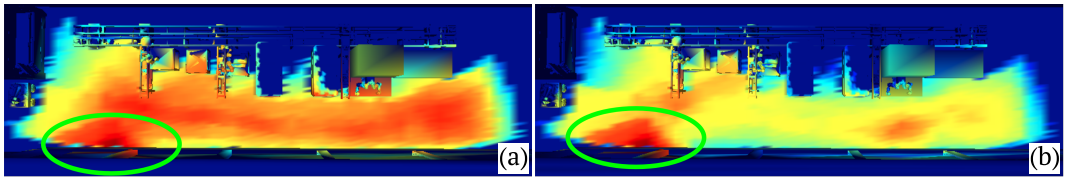


Fig. 5.2 Qualitative results of an unsupervised single-subject experiment. (a) *Dynamic* module disabled. (b) *Dynamic* module active. Without the *Dynamic* module, the resulting VSA is definitely noisier. The target object was an advertising board, circled in green.

Table 5.1 The accuracy in unsupervised experiments.

Method	**CMC Single-Person**				**CMC Multi-Person**			
	Rank-1	**Rank-2**	**Rank-3**	**Rank-4**	**Rank-1**	**Rank-2**	**Rank-3**	**Rank-4**
Vanilla	30.8	46.2	61.5	76.9	25.0	41.7	50.0	66.7
Dynamic	38.5	61.5	84.6	100.0	25.0	50.0	75.0	83.3
Social	-	-	-	-	33.3	58.3	75.0	100.0
SD-VSA	-	-	-	-	33.3	58.3	83.3	100.0

rankings. This curve defines the probability of a specific object having the highest score and, hence, being the correct object within the first N points of the curve. We first analyze the unsupervised single-person sequences. Considering the presence of a single person, we evaluated the system in two configurations: one with the *Dynamic* module and one without (*Vanilla*). Numerically, the *Dynamic* module gives an accuracy boost, and Figure 5.2 shows that VSA maps generated without it are definitely noisier, showing a marked trace on the floor where the subject has passed.

In the unsupervised multi-person sequences, together with the *Vanilla* and *Dynamic* modules, we also consider whether the accumulation on the maps is stopped by the occurrence of social interactions, activating the *Social* module. This leads to four possible combinations (see Table 5.1). As a metric, we report the rank 1-4 accuracies, which show the predominance of the complete *SD-VSA* model as expected.

The single-person unsupervised experiments results are presented in Table 5.1. Since the subjects have never seen the environment, they usually start by navigating the whole hallway before focusing on some specific industrial station. The accuracy of the system solely relying on people's location is low, achieving below 50% at rank-2, which is improved to over 60% introducing the *Dynamic* module. The significant difference in accuracy between rank-1 and rank-2 can be explained by the fact that the machines are distributed very close to each other, indicating that

the correct object is often spatially close to the estimated one. Moreover, since our pipeline is modular, replacing the Gaze Estimation module with a more accurate one could certainly boost the performance described by the CMC curve.

Multi-person unsupervised experiments also present a non-trivial task to solve, as the presence of two individuals can introduce noise in the VSA estimation due to social interactions. Incorporating a SSP-based module helps improve the system's performance. It is important to note that the results of multi-person experiments are generally lower than those of single-person ones since the HPE pipeline faces challenges with occlusion and identity swapping caused by multiple individuals exploring the large scene. Finally, the actual interest of the subjects, collected through the questionnaires, shows that solely relying on Gaze Estimation is not enough, and insights from SSP increase the chance of capturing the actual interest.

Supervised Experiments

The supervised experiments aim to check whether the VSA maps can highlight the object in the scene that the subject was assigned to. In this case, the ground truth is, therefore, the assigned station. As an accuracy measure, we consider a success whether the peak of the VSA map is inside the bounding box surrounding the target object, zero otherwise. The results of these experiments are presented in Table 5.2. As a qualitative result, 5.3 shows the VSA maps after a long social interaction has occurred in the hallway. Without the *Social* module, there are incorrect attention peaks: on the wall behind subject 1 (Figure 5.3(c)) and on the machine behind subject 2 (Figure 5.3(a)). Instead, with *SD-VSA*, we can recognize this interaction and suspend the attention weighting, focusing only on the actual attention of the subjects.

In these experiments, the subjects are now more familiar with the environment. Additionally, since we had assigned the target industrial machine beforehand, subjects naturally performed less to no room exploration, heading straight for the given machine. This explains why the supervised experiments achieve higher results than the unsupervised ones. The results shown in Table 5.2 clearly reflect this difference between the two sets of experiments, providing further evidence of the usefulness of considering SSP insights.

Table 5.2 The accuracy in supervised experiments. In this case, the value tells us if we could identify the specific station assigned to subjects.

Method	Single-Person	Multi-Person
Vanilla	56.3	50.0
Dynamic	62.5	54.2
Social	-	62.5
SD-VSA	-	70.8

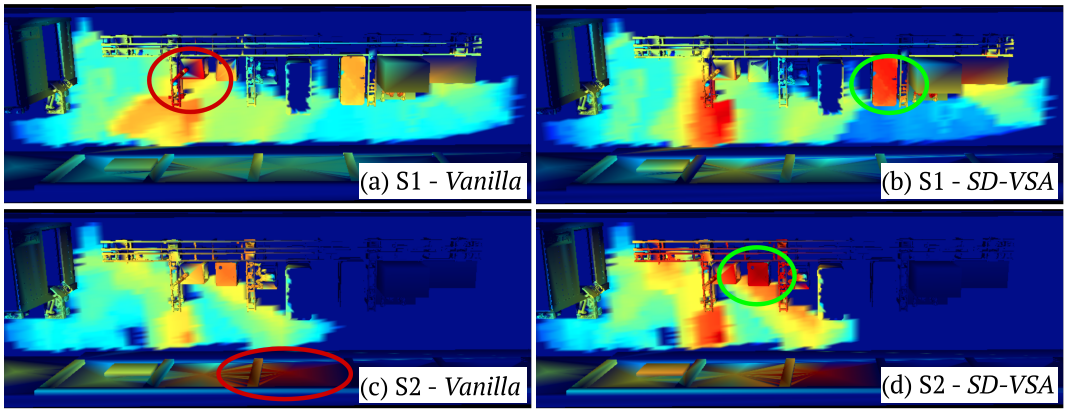


Fig. 5.3 Qualitative results of a supervised experiment with a pair of subjects (S1 and S2) that had a dialogue in the center of the hallway. In (b) and (d), the objects circled in green correspond to the correct most attractive points for those VSA maps. (a) and (c) show instead the false positive targets (circled in red) obtained by disabling the *Dynamic* and *Social* modules.

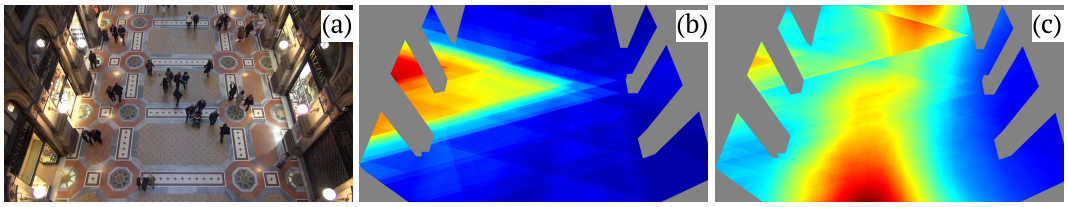


Fig. 5.4 (a) A frame from the GVEII dataset. (b) *SD-VSA* baseline. (c) *Vanilla* baseline: the *Dynamic* and *Social* modules are switched off.

Does the *SD-VSA* Scale Up to Crowded Scenarios?

To prove that the SSP modules of the *SD-VSA* baseline are effective and generalize to highly crowded scenes, we use the GVEII benchmark [11], where hundreds of people walk in an outdoor mall. Positions and group formations are already available as annotations. We compute the VSA with *SD-VSA* and the *Vanilla* baseline, using a simple 3D model and projecting the frustums on the floor for visualization purposes.

Note that here, the view frustums are not computed through 3D head pose estimation; instead, we use the oriented velocity vector derived from the ground truth trajectory data. In Figure 5.4(b), it is possible to note a clearly interpretable global VSA, communicating that the people were interested in the two shop windows on the left, while the bottom-right window shop, which is closed, gather the least attention. In Figure 5.4(c), the results are incorrect since the most interesting part of the scene appears to be the pathway due to the vertical flow of the people.

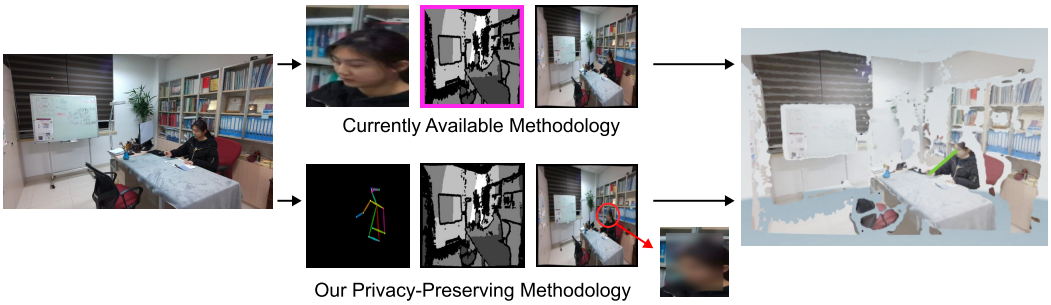


Fig. 5.5 We introduce a novel way to tackle the 3D Gaze Estimation and 3D GTD problems that, differently from the previous approaches, does not use the head crops of the observed people but rather employs upper-body skeletons, depth maps, and scene images where the head is blurred out. The new pipeline not only proved successful in tackling both tasks but offers a concrete way to preserve the identity of the involved people. The currently available state-of-the-art method already used depth maps (in magenta) in its pipeline, so we are not introducing additional information.

5.4 Privacy-Preserving 3D Gaze Target Detection

In this work, we propose a methodology that not only solves the 3D Gaze Estimation task in a novel way, employing estimated upper-body skeletons and ground truth depth maps but also provides a solution for both 2D and 3D GTD while simultaneously preserving the identity of the observed subjects. Inspired by the preliminary work carried out in [81], we build upon the framework and the dataset proposed by [97], replacing their Gaze Estimation module based on head crops with a skeleton and depth-based module, limiting the treatment of scene and face images only to the HPE pre-processing phase, after which all these sensitive information can be discarded (see Figure 5.5). A single attention layer, applied to the combination of upper-body pose features normalized on one anchor joint and convolutional features extracted from the scene depth maps, provided us with accurate 3D gaze vectors. These vectors are then fed to a 3D GTD pipeline, which uses scene images where the people's faces have been blurred out. The proposed method can be trained either in a multi-stage fashion, pretraining the Gaze Estimation module, or in an end-to-end manner. This approach not only achieves state-of-the-art results on the GFIE dataset [97], the most comprehensive dataset available for 3D scenarios, but also represents a novel approach to privacy preservation in the context of 3D GTD.

The main contributions of our work can be summarized as follows:

- We present a novel module for 3D Gaze Estimation that utilizes upper-body skeleton data and the depth map of the scene. This approach achieves outstanding results, showcasing the richness of pose data in understanding a person's gaze direction in 3D.
- We incorporate the above-mentioned new module into a pipeline for 3D GTD, achieving state-of-the-art results on the GFIE dataset [97], the most comprehensive dataset currently available for the task at hand.
- For the first time, we focus on the privacy-preserving capabilities of a vision-based Gaze Estimation and GTD approach in the 3D space, discarding information about the person's facial appearance and concentrating solely on their upper-body poses. By doing so, our approach eliminates the need to use a head backbone pretrained on a large dataset of head orientations.

5.4.1 Methodology

In this section, we introduce our upper-body and depth-based Gaze Estimation module and explain how it enables us to execute the entire 3D GTD pipeline while preserving the privacy of the involved subjects. The proposed architecture is presented in Figure 5.6 and described in Section 5.4.1.

Section 5.4.1 describes our novel Gaze Estimation method and how it fits inside the GTD pipeline. Privacy preservation and how it is enforced is discussed in Section 5.4.1, while in Section 5.4.1 and Section 5.4.1 we provide training and implementation details to enable appropriate reproducibility of the proposed approach.

Gaze Target Detection Pipeline

Given an image or video dataset which offers annotations about the ground truth gaze vector and the 2D/3D gaze targets, the task of 3D GTD consists of finding the 3D point $t^{3D} = (t_x^{3D}, t_y^{3D}, t_z^{3D})$ looked at by the subject, or subjects if extended to social gaze prediction, in the scene and their gaze direction represented by the normalized vector $g = (g_x, g_y, g_z)$. The input data for each gaze prediction includes the image scene s, the aligned depth map d with associated camera parameters k, a binary mask for the face location m, the 2D (e^{2D}) and 3D (e^{3D}) annotations of the point between the eyes and the bounding box (box_h) around the head. Note that the

head bounding box is not employed as a modality but is used to blur the individual's face in the scene image; For further details, see Section 5.4.1.

We adopted the four-module pipeline presented in [97] to extract the 2D and the 3D gaze targets from the 2D image and the 3D point cloud, respectively. The Gaze Estimation module, Figure 5.6(a), computes the normalized three-dimensional gaze vector $\hat{g} = (\hat{g}_x, \hat{g}_y, \hat{g}_z)$, starting from the annotated 3D eye position $e^{3D} = (e^{3D}_x, e^{3D}_y, e^{3D}_z)$ and describing the direction of where the person in the scene is looking. The second module (Figure 5.6(b)) creates a heatmap V defined as the dot product between the unprojected depth map $d' - e^{3D}$, corresponding to the point cloud of the scene centered in the 3D eye position e^{3D}, and the estimated gaze vector $\hat{g}$, with the maximum values representing the points in the point cloud with the highest probability of being looked at. Intuitively, each unprojected (and translated) point represents a vector from e^{3D} to that point, and the heatmap role is to highlight the closest vectors to the estimated gaze direction. A second heatmap $\hat{V}$ is computed as $\hat{V} = ReLU(V)^{\alpha}$, with $\alpha = 3$, to further highlight the most interesting points in the scene. The scene image s, the heatmaps V and $\hat{V}$, and the head location mask m are then concatenated and fed to the third module (Figure 5.6(c)), an encoder-decoder network, that generates the 2D gaze heatmap H used to extract the 2D gaze target.

The last module, depicted in Figure 5.6(d), is used only in the evaluation phase and is responsible for extracting the 3D gaze target from the original point cloud. After extracting the 2D gaze target from the predicted heatmap H as $t^{2D} = \texttt{argmax}(H)$, a 3D volume centered in that point is built in the unprojected (and translated to the e^{3D} position) vector space to enclose a portion of the available points. The vector with the highest similarity to the predicted gaze vector inside the volume is selected as the new predicted gaze vector $\hat{g}$, and the corresponding point is selected as the 3D gaze target from the original point cloud. Below, we discuss the critical ingredients and the processing of our approach and, as a result, the novelty of the pipeline in detail.

The Skeleton-based Gaze Estimation Module

To effectively obtain a pipeline able to preserve the identity of the subjects and avoid using image data related to the appearance of the face, the first step we adopt consists of replacing the original Gaze Estimation method based on face crops [97]. We, therefore, introduce a novel module for Gaze Estimation, *i.e.*, the prediction of the gaze direction in the 3D space that exploits the upper-body skeleton and the depth

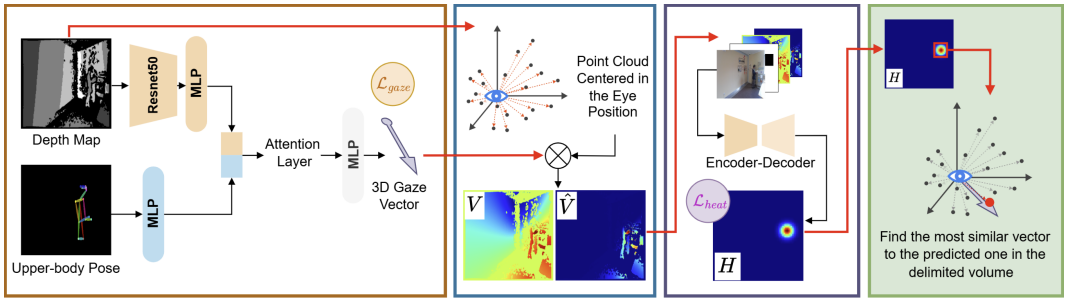


Fig. 5.6 Overview of the proposed approach. (a) The 3D Gaze Estimation module predicts a gaze vector by exploiting the upper-body pose coordinates, processed by a simple MLP, and the convolutional features extracted from the depth map by a ResNet50 [87]. A single multi-head attention layer is applied to the concatenated features before the final MLP to regress the three vector coordinates. (b) The perception module converts the depth map to the unprojected point cloud, from which the 3D eye coordinates are subtracted, and computes two heatmaps highlighting the most interesting part of the scene. (c) The 2D gaze target heatmap is predicted by an encoder-decoder architecture operating on the concatenation of the scene image (with the blurred-out face), the head location mask and the two heatmaps. (d) During the evaluation, the most similar vector to the predicted vector $\hat{g}$ in the unprojected and translated point cloud is selected as the final gaze vector and the corresponding point in the original point cloud as the predicted 3D gaze target.

map as multimodal information to extract a coherent gaze vector. Considering the HPE as a preprocessing step, after which the whole GTD pipeline doesn't require the face appearance, the proposed Gaze Estimation module has proven effective in correctly estimating the gaze direction despite having no access to either the face crop or the scene image. The pose keypoints used in this study are from the hip upwards, for a total of 13 joints, including the arms and wrists, which are extracted using a pretrained top-down pose estimator [219].

The keypoints go through a normalization procedure that uses the neck position and the distance between the neck and hip joints to obtain a scale-invariant representation of the human pose. A simple Multi-Layer Perceptron (MLP) with three layers processes them to expand the features' dimensionality. Concurrently, a convolutional encoder extracts the depth features from the resized depth map. The upper-body and depth features are concatenated, and a single attention layer is applied to select the most helpful features for predicting the gaze direction. Finally, two linear layers are applied to these features to extract the 3D gaze vector $\hat{g}$, which is normalized before being handed over to the next module.

Privacy Preservation

The only step of the proposed pipeline where the face images are used is the HPE preprocessing phase, ensuring that the GTD architecture can offer the claimed privacy-preserving feature. In addition to the use of upper-body poses and depth maps to replace the need to use head crops, the entire scene image used as one of the inputs by the autoencoder module for the prediction of the heatmap H has been edited by applying Gaussian blur in the region defined by the head bounding box provided in the dataset, box_h, hence making sure that no module could use that information to improve the GTD. Examples of edited scene images and skeletons extracted by the HPE method are shown in Figure 5.7.

Training Details

Excluding the HPE phase, where a pretrained 2D pose estimator has been adopted, our GTD pipeline can be trained in two ways: a) multi-stage training, and b) end-to-end manner. For the multi-stage training, firstly, the Gaze Estimation module is trained standalone using only a cosine loss on the estimated gaze vector $\mathscr{L}_{gaze} = 1 - CosineSimilarity(g, \hat{g})$ on the same train-valid-test split as the complete architecture. The corresponding weights are then loaded into the same module, but this time as part of the full pipeline. The latter is then trained in an end-to-end fashion and supervised on the gaze vector estimation with a cosine loss $\mathscr{L}_{gaze}$ and on the 2D heatmap H creation with an L2 loss $\mathscr{L}_{heat}$. The total loss is then computed as $\mathscr{L} = w_{heat}\mathscr{L}_{heat} + w_{gaze}\mathscr{L}_{gaze}$, where w_{heat} and w_{gaze} are two loss-specific weights defined as hyperparameters. Specifically, w_{heat} has a value of 10000, and w_{gaze} has a value of 10, similar to [97]'s implementation. The second training approach, the so-called end-to-end, involves training the entire GTD pipeline, including the Gaze Estimation module, in a unified manner. This method does not rely on any pretrained Gaze Estimation module from either a different dataset or the training split of the dataset used.

Implementation Details

The proposed approach modules are implemented in PyTorch and trained on a machine equipped with an Intel-i9 CPU and a single NVIDIA RTX 4090.

Gaze Estimation Module. The module that manages the estimation of the gaze direction processes depth maps resized to 224×224 pixels using a ResNet50[87] and three linear layers with Rectified Linear Activation Unit (ReLU) and Dropout. A three-layer MLP with ReLU activation function handles the 13×2 upper-body pose data estimated from a pretrained HRFormer [219] and normalized on the neck position and the distance between the neck and hip joints. The attention layer is implemented as a Transformer Encoder Layer with four heads and a feature size of 256. In the multi-stage training regime, the Gaze Estimation module is trained standalone with a batch size of 128 for 30 epochs, the Adam [108] optimizer and a learning rate and weight decay of $1e-4$. Given that prior work employs auxiliary large Gaze Estimation datasets for pretraining their Gaze Estimation module, we take a different approach for our multi-stage training. Instead of using an additional dataset, we utilize the weights obtained from training our Gaze Estimation module solely on the training set of the GFIE dataset [97] with the gaze loss $\mathscr{L}_{gaze}$ (corresponding results are given in Section 5.4.2) to initialize the module within the complete GTD pipeline.

Gaze Target Detection Pipeline. The full pipeline is trained for 50 epochs, with the Adam [108] optimizer, a batch size of 32, and a learning rate and weight decay of $1e-4$. The images and depth maps are resized to 224×224 pixels. Random crop, flip, saturation, brightness and contrast adjustments have been applied as augmentations. The predicted perception heatmaps V and $\hat{V}$ have a size of 224×224 pixels, while the final 2D target detection heatmap H has a size of 64×64 pixels. The ground truth heatmap is generated by a Gaussian centered in the gaze target [160]. The encoder is implemented as a ResNet50 [87], pretrained on Imagenet [53], while the decoder consists of two convolutional layers and three deconvolutional layers.

5.4.2 Evaluation

In this section, we describe the dataset used (Section 5.4.2) for our study, detailing the evaluation setup along with the 2D and 3D metrics used for Gaze Estimation and GTD (Section 5.4.2). We then showcase the results of our Gaze Estimation module (Section 5.4.2), comparing them to the most related work [97]. Next, we report the results of 2D/3D GTD (Table 5.3), corresponding to the entire pipeline and comparing our findings against state-of-the-art methods. Finally, we present the qualitative results (Section 5.4.2) and ablation study (Section 5.4.2).

Dataset

Datasets annotated for 3D Gaze Estimation and 2D/3D GTD are particularly scarce. Consequently, we use the largest available dataset of this kind, GFIE [97], to evaluate our proposed method. This dataset allows us to compare our results with state-of-the-art methods and conduct an ablation study to examine the impact of using only a portion of the skeleton and the effect of blurring the face.

The GFIE dataset [97] was captured by using a laser rangefinder and an Azure Kinect RGB-D camera mounted on a stable platform. The laser rangefinder projects a laser spot to guide the subject's gaze, ensuring continuous fixation during recording. This way, the distance measurements from the laser rangefinder can be logged. Annotations encompass head bounding boxes, 2D/3D gaze targets, and 2D/3D eye locations. The annotations were performed semi-automatically. For example, a face detection method [224] was used, followed by manual verification. For 2D gaze targets, after detecting laser spots in RGB images, a team of annotators validates the accuracy of these detections. Using the distance measurements obtained from the laser rangefinder, the positions of the gaze targets in 2D are mapped into corresponding 3D coordinates in the real-world space. Additionally, using a facial landmark detector, the locations of left and right eye landmarks were identified within cropped head images. The 2D eye position corresponds to the midpoint between these landmarks, while the 3D eye position was computed by unprojecting from the depth map's facial landmarks. This dataset includes diverse gaze behaviours from 61 subjects (27 male and 34 female), covering various activities. It consists of a total of 71,799 frames. The dataset is divided into a training set containing 59,217 frames, a test set with 6,281 frames, and a validation set containing 6,281 frames. Notably, the subjects and scenes in the training set were excluded from the test and validation sets. To validate our approach, we use the same evaluation protocol defined in [97], and during training, we use the annotations for the 2D and 3D gaze targets and the annotated 3D gaze vector as the ground truth.

Evaluation Metrics

While our primary focus remains on the 3D evaluation metrics and the 3D scene in general, we also report the results for the 2D metrics, usually adopted for the standard GTD task. The standard 3D metrics involve distance, and the Angle Error is defined as follows. *3D Distance* measures the Euclidean distance between the

predicted and actual gaze points in a 3D space. The unit is meters. *Angle Error* quantifies the angle difference between the predicted and actual gaze directions. It is computed as the `arccos` of the cosine similarity between the normalized ground truth and predicted gaze vectors. The unit is degrees. On the other hand, 2D metrics include the Area Under the Curve (AUC) and the 2D Distance, described as follows. *AUC* utilizes the predicted heatmap to plot the Receiver-Operating Characteristic (ROC) curve, as proposed by [104]. *2D Distance* computes the Euclidean distance between the predicted and actual gaze points, using the `argmax` of the predicted 2D heatmap, assuming an image size of 1×1. While the Angle Error is a metric for 3D Gaze Estimation, all other metrics pertain to 2D/3D GTD tasks.

Results of the Gaze Estimation Module

Without injecting our proposed module for Gaze Estimation into our complete GTD pipeline, we trained and tested it alone. We performed the same for the face crop-based approach in [97]. Differently from our module, [97] pretrained their network that gets the head crops as input on another Gaze Estimation dataset called Gaze360 [107]. However, we do not use any auxiliary datasets to pretrain our model. [97], while performing very well on the training set with a 3D Angle Error close to 3°, when it comes to the testing set, it achieves an Angle Error of 19.7°, showing an overfitting characteristic. On the other hand, our training and testing performance is consistent, and we perform a 3D Angle Error of 16.5°, which is not only remarkably better but also overfitting-free.

Results on the GFIE Dataset

The results obtained by running our full pipeline on the GFIE dataset [97] are reported in Table 5.3. As summarised below, we compare our method with several approaches for detecting gaze targets in both 2D and 3D.

- *Random* approach refers to gaze targets chosen randomly within the image for 2D and within the point cloud for 3D space.
- *Center* approach positions the gaze targets at the center of the image for 2D space and at the center of the point cloud for 3D space.
- For the existing methods that focus on GTD in 2D, such as *GazeFollow* [160], *Lian et al.* [121], and *Chong et al.* [44], they have been extended to estimate

3D gaze targets by unprojecting the 2D gaze targets into 3D space in [97]. Additionally, *RT-Gene* [60] and *Gaze360* [107], which estimate 3D gaze directions in unconstrained environments, have been adapted to the same evaluation protocol in [97]. The most similar vector to the predicted gaze vector within the unregistered depth map has been selected as the actual gaze vector, and the point obtained by adding this vector to the eye coordinates is considered the 3D gaze target.

- *Hu et al.* [97] is the most relevant work, which uses head crops, scene images (without blurring), and depth maps to build a point cloud and simultaneously estimate the gaze direction and location in both 2D and 3D.

The baseline results, obtained with the *Random* and *Center* heuristics, do not offer particular insights about the problem, achieving comprehensively high error values in both scenarios. The methods built explicitly for Gaze Estimation (*Rt-Gene* [60] and *Gaze360* [107]) perform well in estimating a suitable gaze vector, while obtaining poor results on the gaze target selection.

Our methodology, trained with a multi-stage strategy or an end-to-end strategy and based on a Gaze Estimation method that exploits the attention mechanism on the feature extracted from both upper-body poses and depth maps and operates on scene images with the faces blurred out, outperforms all state-of-the-art in three metrics out of four. More importantly, both the 3D metrics have been improved with respect to [97], with a gain of 27 millimeters on the 3D distance and a 1.8° improvement on the 3D Angle Error achieved by the multi-stage strategy, with the Gaze Estimation module trained alone before the full pipeline training. Improving Gaze Estimation is, therefore, of fundamental importance in solving the task of GTD in 3D, even more than in 2D scenarios. Since most of the models' inputs are sensory-acquired 3D data (depth maps, camera parameters, and eye positions), a better estimation of the gaze vector would allow focusing on a relatively small number of potential 3D targets.

Qualitative Results

In Figure 5.7, we present some qualitative results obtained on the GFIE dataset [97]. As it can be seen from the analyzed samples, our method can estimate 3D gaze vectors and gaze targets (in blue) very close to the ground truth annotated ones (in red). More interestingly, our Gaze Estimation module seems to operate well regardless of the orientation of the person, with upper-body skeletons front facing

Table 5.3 Results on the GFIE dataset [97] for the 2D and the 3D scenarios. The results of the prior work are taken from [97]. *Dist.* stands for Distance. The **bold** numbers represent the **best** results. The second-bests are underlined. H, D, UP, and S stand for *Head crop*, *Depth map*, *Upper-body Pose*, and *Scene*, respectively.

Method	Modalities	Privacy	3D Metrics		2D Metrics	
			Dist. ↓	Angle Error ↓	AUC ↑	Dist. ↓
Random	S	✗	2.930	84.4°	0.585	0.425
Center	S	✗	2.510	87.2°	0.614	0.287
GazeFollow [160]	H, S	✗	0.856	41.5°	0.941	0.131
Lian et al. [121]	H, S	✗	0.542	26.7°	0.962	0.091
Chong et al. [44]	H, S	✗	0.455	20.8°	0.972	0.069
Rt-Gene [60]	H, S	✗	0.552	21.0°	0.823	0.123
Gaze360 [107]	H, S	✗	0.540	19.8°	0.821	0.130
Hu et al. [97]	H, S, D	✗	0.311	17.7°	0.965	**0.065**
Ours (Multi-stage)	UP, D, S (blurred)	✓	**0.284**	**15.9°**	**0.983**	0.083
Ours (End-to-end)	UP, D, S (blurred)	✓	0.292	16.2°	0.973	0.083

Table 5.4 Ablation study in terms of modality and the type of input scene image (*i.e.*, face blurred and not blurred) on the GFIE dataset [97]. *Dist.* stands for Distance. Note that depth maps and scene images are fundamental and must be included for the 3D GTD to function properly; thus, their removal cannot be ablated.

	Head Crop	Blurred Face	Full Body	Upper Body	3D Metrics		2D Metrics	
					Dist. ↓	Angle Error ↓	AUC ↑	Dist. ↓
	✓	✗	✗	✗	0.311	17.7°	0.965	0.065
	✗	✗	✗	✓	0.314	16.1°	0.981	0.086
	✗	✓	✓	✗	0.310	16.7°	0.979	**0.059**
Multi-stage	✗	✓	✗	✓	**0.284**	**15.9°**	**0.983**	0.083
End-to-end	✗	✓	✗	✓	0.292	16.2°	0.973	0.083

the camera (fourth row), on the side (fifth row), or captured from the back (first row). Additionally, in the third row, a subject interacts with a robot, both looking and touching it, demonstrating the relationship between body position, the performed action, and, consequently, where the gaze is directed.

Ablation Study

We report ablation results (Table 5.4) performed by removing or changing the used modalities for the proposed method. We also present the results of the proposed method when trained using either the end-to-end approach or the multi-stage training process. As shown, using the upper-body skeleton instead of the complete 17-joint skeleton is slightly beneficial for the 3D Distance compared to the baseline (third



Fig. 5.7 Qualitative results of our method on the GFIE dataset [97]. Each row represents a single sample. (First Column) We use resized depth maps as input for our Gaze Estimation module. (Second Column) The estimated upper-body skeletons are visualized on a black background before being normalized. (Third Column) Scene images where the faces are blurred out before being used by the encoder-decoder module to predict the 2D gaze target heatmap. (Fourth Column) The final point cloud of the scene, with the ground truth gaze vector in red and the estimated gaze vector in blue.

row). Interestingly, using the entire skeleton leads to the best 2D Distance score. This can happen due to the optimization process of the two losses, $\mathscr{L}_{gaze}$ and $\mathscr{L}_{heat}$. In

this case, it seems the latter was favoured over the former. The second row shows that using the model without blurring the faces benefits the 2D AUC metric. The pipeline trained with a multi-stage policy, firstly training the Gaze Estimation module alone and loading its weights before launching the training on the full pipeline, achieved the best results on both the 3D metrics, with a minimum of 15.9° for the Angle Error. However, if a faster training procedure is required, the end-to-end strategy still provides solid results and is preferable in such a situation.

5.5 Discovering the Gazed Target in a Shopping Mall

In this project, conducted in collaboration with other universities, we aim to verify how computer vision and deep learning techniques can support individual profiling. Although in our research we tested our framework on a real-world application as a reference, *i.e.*, the analysis of customer behaviour inside a shopping mall to provide personalized suggestions at digital signage terminals, our primary goal is to investigate key scientific computer vision issues that are central to such and similar applications and develop innovative solutions w.r.t. the state-of-the-art. This application may, therefore, be interesting from an industrial point of view to verify that Gaze Estimation and GTD methodologies are indeed applicable in the real world, particularly with adequate performance that allows for interfacing with a whole range of services. Our focus in the project is the development of a Gaze Estimation-based methodology to track the gaze of customers across different shop windows, gathering real-time preferences over one product rather than another. Before describing the actual method, it can be useful to have an overview of the topics and challenges we addressed within the project. We focused on the following research topics:

- Re-identification of individuals, by creating anonymized identities in open-world settings to allow continuous re-identification of people across different locations.
- Tracking and behaviour analysis, by understanding interests and detecting individuals' actions alone or in a group.
- Extraction of personal stable traits, by learning hidden information such as social class and personality from clothing and behaviours.
- Extraction of personal temporary feelings, by understanding emotional status and attention from face analysis. To foster the usage in real contexts, we designed privacy-respectful solutions.

In this scenario, we tackle a special case of GTD in which the interest targets are always outside the video frame. While in most cases the problem is tackled in the 2D image space [44, 160, 198], as discussed in Section 5.2, we decide again to extend our approach to work in a fully 3D manner. The simulated shopping hall, shown in Figure 5.8, consists of three shop windows, with a camera placed inside each one to record the outside. When a customer approaches one of the windows, a face crop is extracted and sent to the re-identification module to track the person's identity.

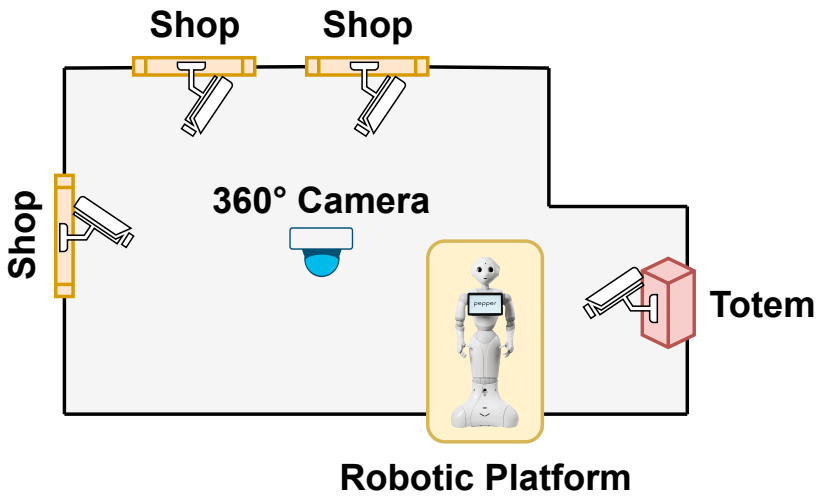


Fig. 5.8 Setup of the **I-MALL** demo.

Traditional face recognition modules assume a closed-set scenario with probe images containing identities enrolled in the gallery. A more realistic scenario is open-set, where probes may contain subjects not enrolled in the gallery, and the recognition system must be able to detect and reject such probes. A more challenging scenario is open-world, where identities are learned incrementally and without supervision in the gallery as soon as they are observed. Parametric learning methods like deep neural networks are natively designed to perform closed-set recognition and have been adapted in a few cases to the open-set. Recent research on Neural Turing Machines [74] showed that deep networks may be enhanced by an external memory to quickly integrate information about new items, ensuring that salient but statistically infrequent data are stabilized in the class representation. Such a model does not support learning from video in any way. The **I-MALL** framework we propose uses deep networks to perform detections and extract representations of face observations, exploiting an external memory module to incrementally collect them, and a smart filtering mechanism that assigns observations to identities and decides their relevance to be learned. We use memory to break up the temporal correlation between consecutive instances that disrupt the non-id nature of video data. In this way, identity clusters are incrementally built, putting together frequent and rare observations without reference to their temporal occurrence. Since the individual outfit is used to learn personality, we exploit outfit features to improve the purity of the clusters. Continuity of face and outfit appearance in consecutive frames

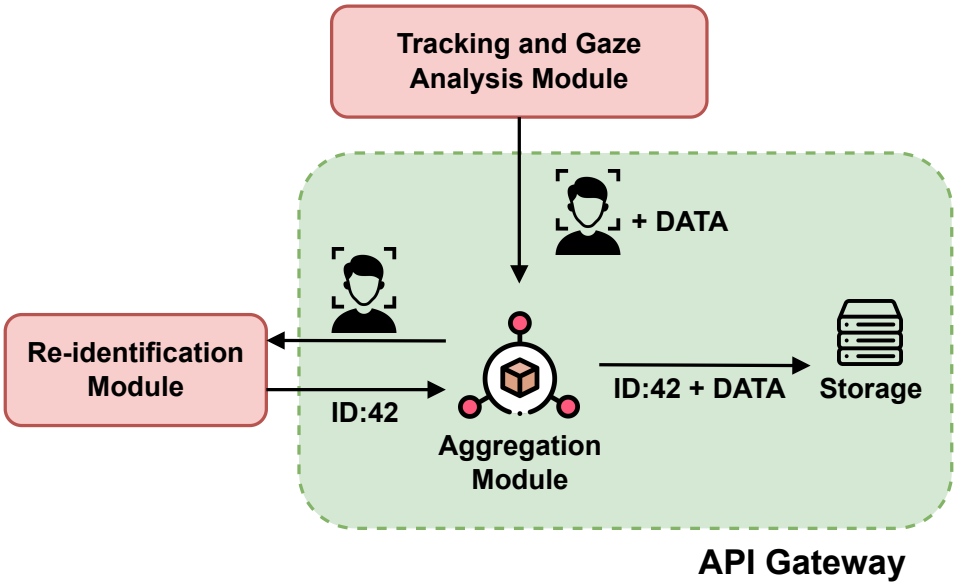


Fig. 5.9 The data aggregation module functional schema.

is used as a form of self-supervision to decide the instantiation of a new identity. The data aggregation module, responsible for the fusion of data coming from the computations of all the other functional modules, is shown in Figure 5.9.

If the system recognizes the same person for a sufficient number of consecutive frames, a pretrained Gaze Estimation model [226] is activated, which estimates the 3D gaze vector of a client towards the shop window. As long as the person stands sufficiently still and faces the window, an attention value is summed over time above the intersection between the gaze vector and the surface corresponding to the shop window. Each showcase can be divided into N tiles, respectively associated with N possible objects of interest within it. In our case, each window is divided into three sections of equal size. At the end of a visit session, the point with the highest attention value will indicate the section, and thus the object, of greatest interest, which will be updated with the corresponding ID within the user's profile. In addition to the single most interesting object, it is also possible to obtain a ranking of interest for each visitor of all the objects (of the mall or the shops individually). Product ranking, synchronized in real time with a central database, is managed through an API-based infrastructure, which allows data to be collected and interfaced from all intelligent modules and services within the shopping mall. These include a MIVIABot (based on the Pepper robotic platform), *i.e.*, a social robot in charge of interacting in a smart way with the human, suggesting the specific product they may

be interested in depending on its biometric characteristics, and a totem to display a multimedia application that can re-identify, through a dedicated camera, the user and retrieve the profiling data stored by the system in its knowledge-base.

5.6 Findings

In this chapter, we proposed three different approaches to analyze the gaze of one or multiple people, thus trying to understand the intrinsic motivation behind people's movements (the *why*). In fact, the direction of the gaze and its target can give us an idea of the intentions and possible future actions that the observed person will take.

In Section 5.3, we proposed **SCENE-pathy**, a benchmark dataset to study the VSA of people towards a complex 3D scene. Along with the RGB video streams, we provide complete annotations of the involved subjects' 3D position and gaze direction, extracted from Hololens 2 devices, and questionnaires reporting the most interesting areas they visited. The proposed baselines, created joining HPE techniques and SSP insights, allowed us to demonstrate that more accurate VSA maps can be obtained by considering the subjects' movement speed and the possible occurrence of social interactions in multi-person scenarios. The main limitation of this work is the difficulty in extending the proposed methodology to very different environments and scenes, requiring an ad-hoc study for each new environment: the construction of the 3D model alone requires a not irrelevant effort. In addition, when people move too far away from the camera, the information available about their body pose and head orientation is much more inaccurate, cascading damage to the mapping of the VSA on the scene. As possible future directions, there is an in-depth study of methodologies for Gaze Estimation on low-resolution images and videos and the analysis methodologies for automatic estimation of the scene model, which could help generalize to more complex environments.

The GTD task, often referred to as gaze-following, is a fundamental proxy task for multiple vision problems regarding the behaviour and interaction of people between themselves and with the environment. Understanding where and what a person is observing can give us hints about their current interests and future actions, whether they are interactions with humans, with objects in the environment, and, more recently, with collaborative robots that are increasingly present in public and workplaces. In Section 5.4, we proposed a novel way to tackle 3D Gaze Estimation and 3D GTD by focusing on the upper-body pose and the scene depth map, outperforming the previous state-of-the-art. The combination of the skeleton-based Gaze Estimation module and the blurring applied on the face in the scene image allowed us to present an effective method that can also preserve the person's sensitive data, such as the face appearance, generally used to solve these types of tasks. When pose estimation fails, our upper-body-based gaze estimation and

target detection pipeline might not perform as effectively. However, this issue is also present in state-of-the-art methods primarily relying on head crops. For the dataset used in this paper, we manually reviewed the body poses for a subset of the testing data and did not observe any failures with the pose estimation method employed. Currently, there is a lack of multi-person 3D GTD datasets. Consequently, our work is constrained by the complexity of scenes and the availability of a dataset that only features single individuals per scene. Unfortunately, our **SCENE-pathy** dataset does not currently suit to be used as a benchmark for GTD, offering only one environment, lacking precise annotation on gaze targets, and including only people wearing Hololens 2, which changes their appearance, especially when observed from a distance. Furthermore, an interesting future direction is the inclusion of activity recognition as part of the pipeline, exploiting the gaze direction and target as a proxy for explicitly predicting the future intentions of one or more people.

Finally, in Section 5.5, we applied our findings on gaze analysis to a real industrial case, estimating customers' product preferences in a simulated shopping mall using the direction of their gaze and combining the results from different shop windows through an API-based software infrastructure. We believe the project's results are general enough to find application in many different real contexts beyond those explored in the experiments. The trend towards environment sensorization and personalized on-demand services will, even more, enhance their relevance in the future.

Chapter 6

Conclusion

In this final chapter, we discuss the findings and insights presented in this thesis, reviewing the contributions made in each area of human behaviour understanding from a computer vision point of view. For each topic, i.e., Human Trajectory Prediction (HTP) in indoor environments, Human Pose Estimation (HPE) and Human Pose Forecasting (HPF) from the robot's perspective and Gaze Estimation with Gaze Target Detection (GTD) in 3D scenes, we review the datasets and methods proposed, highlighting their results and scientific achievements. We also account for the limitations and the open issues we faced in developing these solutions, which are themselves as important as the successes. After that, we present some possible future works and chances for improvements, hoping that they can help the research community in creating novel and more sophisticated computer vision techniques to serve the digital world of tomorrow.

In the thesis introduction, in Chapter 1, we talked about how more and more areas of our world are transforming due to the deployment of new technologies dedicated to studying human behaviour. This thesis aims to consider some of the fundamental aspects that make up human behaviour, in particular when considering the human movement w.r.t. the environment and other agents acting in it, and analyse them from the perspective of the computer vision and deep learning techniques available today. Moreover, our goal is to tackle more specific aspects related to each of the tasks, to offer, on the one hand, new data collections for novel scenarios and, on the other hand, to develop innovative and dedicated solutions to demonstrate that these scenarios are, in fact, addressable. Innovative because they not only allowed us to obtain outstanding results when compared to the state-of-the-art but also, and especially, because they allowed us to take into consideration aspects usually left out in academic research, such as respect for the privacy of the observed people and the need to ensure safety in the use of these systems (*e.g.*, when tackling Human-Robot Interaction (HRI) scenarios).

We took into account three different domains related to human behaviour understanding, all of them focusing on analysing images or videos to extract information about human movement and intentions. The first one, covered in Chapter 3, is Human Trajectory Prediction (HTP) in indoor environments, which includes motion trajectories that are generally more difficult than the better-known and considered outdoor trajectories. In Chapter 4, we estimated human poses and their evolution in a scenario in which the visibility of the observed people is more limited than ever, having as input only the video streams from the on-board cameras of a quadruped robot. Finally, we presented three different projects related to 3D gaze analysis in Chapter 5, proposing a new dataset for mapping visual attention directly on a 3D scene, a novel methodology for Gaze Target Detection (GTD) with privacy preservation and a real-world application of Gaze Estimation in a shopping mall.

After considering these three research domains and after outlining their interconnections and possible solutions, we now have the opportunity to analyse human activity, particularly motor activity, in its entirety, answering the questions: *where* will the observed people move, *how* will they move their body to get there and *why* will they move there, what is their focus of attention? Let us now review the contributions introduced in each of the previous chapters of this thesis.

Human Trajectory Prediction *in Indoor Environments*. In Chapter 3, we presented **SITUATE**, a novel methodology for HTP in indoor environments, exploiting equivariant and invariant feature learning modules and a self-supervised representation learning module to analyse the scene map and its semantics. In addition to the great results obtained in indoor scenes, we achieved comparable performance to other state-of-the-art methods when facing outdoor scenarios. We performed ablative studies on our modules and highlighted their importance in multiple prediction regimes. We believe that indoor HTP, and in general indoor human behaviour understanding, is a fascinating area of research, certainly of utmost interest given the constant development of smart living and smart environments technologies. Indoor computer vision applications can offer new challenges from a scientific perspective and opportunities to create more human-centered techniques.

Human Pose Estimation and Forecasting *from the Robot's Perspective*. In Chapter 4, we presented **HARPER**, the first dataset to study human-robot collision prediction from the Point of View (PoV) of the quadruped robot Spot. **HARPER** consists of 15 actions performed by 17 participants, acting realistic interactions in a collaborative scenario, some of which contain a collision between the human and the robot. This project aimed to offer the possibility of tackling human behaviour understanding from the perspective of a quadruped robot, a fairly complicated task due to the location and quality of the on-board cameras and the quick movements of the robot. We proposed three benchmarks and related baselines, which, solved one after another, constitute a pipeline already capable of predicting some of the collisions caused by humans toward the robot. The first task we tackled was 3D Human Pose Estimation (HPE), *i.e.*, estimating the location of the visible human joints in 3D, followed by the 3D Human Pose Forecasting (HPF) and Collision Prediction (CP) tasks, *i.e.*, predicting the future body joints position and whether this movement can lead to a collision. Equipping one of the most sophisticated robots on the market (Spot by Boston Dynamics) with a system to predict possible collisions with the people around it is a worthwhile endeavour in view of the use of cobots such as these in industrial and assistive settings, where the interaction with other agents is inevitable.

Gaze Estimation and Gaze Target Detection *in the 3D Space*. In Chapter 5, we presented the tasks of Gaze Estimation and GTD, dealing with estimating the

direction and target of the gaze, respectively. These tasks are usually solved by observing one or more persons from a third-person PoV. We proposed three solutions, each based on different assumptions and aimed at investigating the visual focus of attention in a specific scenario. In Section 5.3, we presented **SCENE-pathy**, a dataset for studying the Visual Selective Attention (VSA) towards the 3D scene. This is a special case of GTD, where the goal is not to predict the gaze target of an observed subject at a certain time but to map their attention on a 3D point cloud over time. Moreover, we acquired video streams where one or a couple of people navigate the scene, observing the various interest elements around them and occasionally interacting between them. We collected Hololens 2 gaze and position data to compute a precise ground truth of the most observed parts of the scene. We then developed a baseline based on HPE and head pose estimation techniques that, combined with some Social Signal Processing (SSP) cues, *i.e.*, movement speed and occurrence of social interaction, achieved interesting results.

Following that, we presented a 3D GTD algorithm in Section 5.4. While most GTD methods are deployed in 2D scenarios, our method is designed to estimate the gazed target also in 3D, using as input a depth map of the scene. One of the innovations of this project is a methodology for estimating gaze direction based on the upper-body pose of the person being observed and not their appearance. Hence, our methodology can ensure the privacy preservation of the observed person. We achieved state-of-the-art 3D estimation results on GFIE [97], the most comprehensive and recent dataset for this task, while obtaining comparable results on the 2D GTD task. Finally, in Section 5.5, we demonstrated the real-world applicability of Gaze Estimation techniques like the one discussed above by integrating a 3D GTD framework inside a functioning smart shopping mall demo. This final work has been carried out in collaboration with other research groups, each focusing on a different human behaviour trait and a specific framework component. After being combined with each other via a centralized API infrastructure, these parts can profile the users and offer them an improved customer experience.

6.1 Limitations and Future Works

In this last section, we discuss some of the limitations that emerged in the development of the projects that make up this thesis, highlighting things that could be

changed and what might be some research directions for future work related to the study of human behaviour.

Complex Solutions Require Complex Pipelines. Most of the scenarios addressed in this thesis involve specific applications for the study of human behaviour, *e.g.*, indoor HTP or HPE from a novel and incomplete perspective (*i.e.*, the quadruped robot). Developing solutions for tasks of this type often requires creating complex pipelines that contain multiple deep learning models within them and in which the results of one model are the input for the next. This is the case, for example, of the baseline proposed in **SCENE-pathy**, in which gaze direction is estimated in cascade using a HPE model, a 3D lifting model, and a tracking model to be sure to consider the same person across frames. The same thing is evident in **HARPER** and the methodology presented in Section 5.4, in which HPE is a critical step for proper pipeline operation. Thus, there is a clear need for robust techniques for solving the tasks underlying the most complex pipelines or for end-to-end solutions whose training is spread throughout the estimation pipeline. Future works should focus on developing methodologies that can work with partial and potentially incorrect data to handle cases where any of the pipeline components may fail. Training the downstream architectures with weak supervision can help account for imperfect data. Also, developing end-to-end and Multi-Task Learning (MTL) solutions, *i.e.*, solving multiple tasks at the same time, can help mitigate the impact of one training loss over the others. This obviously opens the way to innumerable challenges worth investigating, such as the transition from manual selection of the weights for each loss (as done in Section 5.4) to techniques for automatically selecting these weights, such as [47].

Furthermore, complex pipelines usually require computation times beyond the ability to run in real time. In this context, the pipeline proposed in **HARPER** needs to be optimized to operate directly on-board, with a major bottleneck caused by the diffusion-based model used to reconstruct the missing poses. It would be worth investigating solutions that can generalise better and do not need to be retrained when new data is obtained.

Specific Applications Require Specific Annotations. For the same reasons described in the previous point, since the problems addressed in this thesis are particular and specific to certain use cases, it is necessary to create novel datasets. Specifically,

accurate annotations on different modalities must be provided, along with the visual data, limiting the need for manual annotations whenever possible. In gaze analysis, for example, it is well known that many annotations used to train GTD models are done manually, which leads to discordant annotations for the same sample and errors when verified qualitatively. This thesis partly attempts to compensate for this lack of datasets, but the proposed collections remain limited by the diversity of the considered scenarios and situations. Future work on these topics should investigate ways to expand the availability of such data. For instance, by leveraging innovative sensors for acquisitions (*e.g.*, the Hololens 2 we used to acquire **SCENE-pathy**), teaming up between researchers to create large shared datasets, and exploring the potential of synthetic, automatically generated data that require no special effort to be annotated. Specifically, this last one is a very interesting topic to investigate, as creating synthetic data for studying human behaviour is particularly complex due to the very nature of such data.

Beyond Traditional Modalities. Although one of the claims made in this thesis is a call for the use of computer vision methodologies that can preserve the privacy of the people being observed, it is clear that in many respects and for many tasks, it is still difficult to imagine abandoning the use of features related to a person's appearance. We are not just referring to features related to people's faces, but all those factors that allow for the re-identification of a person, such as the way they move and the clothes they wear. If a person's face is visible in most frames, with a high enough resolution, then it will probably still be preferable for every gaze analysis, given the possibility of using the eye's information directly. In Section 5.4, we decided to move in a different direction, exploiting the skeleton to infer the gaze direction of the observed people. Future developments could combine these two techniques, allowing the face and eyes to be exploited when visible and the position of the head when the person is not facing the camera or barely visible. In general, given the possibility of having to deal with very complex case studies, such as the presence of people very far away from the camera lens and therefore with very small proportions, it will be necessary to rely on multiple modalities to study human behaviour and interactions, be it depth, 3D scans, sounds, or categorization of common actions. Finally, a very interesting research effort we didn't explore in this thesis (and which is even more like the examples we can find in the real world) is devoted to studying human behaviour from multi-camera setups, where

the main difficulty is associating the different estimates with a single person, using re-identification or tracking techniques.

Reasoning in the Latent Space. In a research work I did not include in this thesis, called *Disentangled Latent Spaces Facilitate Data-Driven Auxiliary Learning* [179], we explored the possibility of discovering novel auxiliary tasks to be used in a MTL framework by disentangling (separating) the generative factors that constitute the input images. Although the results obtained are exciting, we have yet to be able to verify the feasibility of exploiting latent space to improve performance on tasks other than classification. The breakdown of latent spaces is one of the most interesting future works that can arise from the work presented in this thesis, namely the creation and analysis of latent spaces in which the various components of human behaviour can be identified independently of each other so that they can be used to solve different and multiple tasks together. A preliminary example is what has been done in [14], where a disentangled latent space is used to generate more realistic and diverse human body movements. The ability to separate the components of human behaviour can help improve performance in downstream tasks and, more importantly, understand the impact of each of these components, making future deep learning solutions more interpretable (and therefore more manageable). Explainable deep learning models are at the core of a safe and sustainable evolution of this technology, fostering trust, transparency, and accountability in real-world applications.

6.2 Closing Remarks

I would like to conclude this thesis with a personal reflection on the work that has been done and that led to this manuscript. I am convinced that the study of human behaviour is of primary importance and among the most interesting applications of technologies such as neural networks and machine learning in general. However, this is true as long as the research results are used to help people in need, offer valuable services and, in general, create a better society in which to live. Those working on these tools should be responsible for ensuring equality, privacy and safety for the end users and the community.

Acknowledgements

Some of the work included in this thesis was partially supported by the Italian MIUR PRIN 2017, Project Grant 20172BH297: I-MALL - improving the customer experience in stores by intelligent computer vision and within the PNRR research activities of the consortium iNEST (Interconnected North-Est Innovation Ecosystem) funded by the European Union Next-GenerationEU (Piano Nazionale di Ripresa e Resilienza (PNRR) – Missione 4 Componente 2, Investimento 1.5 – D.D. 1058 23/06/2022, ECS_00000043).

References

[1] Henny Admoni and Brian Scassellati. “Social Eye Gaze in Human-Robot Interaction: A Review”. In: *Journal of Human-Robot Interaction* 6.1 (Mar. 2017), p. 25.

[2] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. “A Spatio-temporal Transformer for 3D Human Motion Prediction”. In: *2021 International Conference on 3D Vision (3DV)*. IEEE. IEEE, Dec. 2021, pp. 565–574.

[3] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. “Social LSTM: Human Trajectory Prediction in Crowded Spaces”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016.

[4] Fernando Amodeo, Noé Pérez-Higueras, Luis Merino, and Fernando Caballero. “FROG: A new people detection dataset for knee-high 2D range finders”. In: *arXiv preprint arXiv:2306.08531* (2023).

[5] A ASHRAE. *Guideline 10P, Interactions Affecting the Achievement of Acceptable Indoor Environments*. 2010.

[6] Andrea Avogaro, Andrea Toaiari, Federico Cunico, Xiangmin Xu, Haralambos Dafas, Alessandro Vinciarelli, Emma Li, and Marco Cristani. “Exploring 3D Human Pose Estimation and Forecasting from the Robot’s Perspective: The HARPER Dataset”. In: *arXiv preprint arXiv:2403.14447* (2024).

[7] Görkay Aydemir, Adil Kaan Akan, and Fatma Güney. “ADAPT: Efficient Multi-Agent Trajectory Prediction with Adaptation”. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2023.

[8] Inhwan Bae and Hae-Gon Jeon. “Disentangled Multi-Relational Graph Convolutional Network for Pedestrian Trajectory Prediction”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.2 (May 2021), pp. 911–919.

[9] Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. “Learning Pedestrian Group Representations for Multi-modal Trajectory Prediction”. In: *Computer Vision – ECCV 2022*. Springer. Springer Nature Switzerland, 2022, pp. 270–289.

[10] Inhwan Bae, Young-Jae Park, and Hae-Gon Jeon. “SingularTrajectory: Universal Trajectory Predictor Using Diffusion Model”. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2024, pp. 17890–17901.

[11] Stefania Bandini, Andrea Gorrini, and Giuseppe Vizzari. “Towards an integrated approach to crowd analysis and crowd synthesis: A case study and first results”. In: *Pattern Recognition Letters* 44 (July 2014), pp. 16–29.

[12] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. “BEiT: BERT Pre-Training of Image Transformers”. In: *International Conference on Learning Representations*. 2021.

[13] Jun Bao, Buyu Liu, and Jun Yu. “ESCNet: Gaze Target Detection with the Understanding of 3D Scenes”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 14106–14115.

[14] German Barquero, Sergio Escalera, and Cristina Palmero. “BeLFusion: Latent Diffusion for Behavior-Driven Human Motion Prediction”. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2023, pp. 2317–2327.

[15] Emad Barsoum, John Kender, and Zicheng Liu. “HP-GAN: Probabilistic 3D Human Motion Prediction via GAN”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, June 2018, pp. 1499–149909.

[16] Federico Bartoli, Giuseppe Lisanti, Lorenzo Seidenari, and Alberto Del Bimbo. “User Interest Profiling Using Tracking-free Coarse Gaze Estimation”. In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE. IEEE, Dec. 2016, pp. 1839–1844.

[17] Loris Bazzani, Marco Cristani, Diego Tosato, Michela Farenzena, Giulia Paggetti, Gloria Menegaz, and Vittorio Murino. “Social interactions by visual focus of attention in a three-dimensional environment”. In: *Expert Systems* 30.2 (2013), pp. 115–127.

[18] Federico Becattini, Giuseppe Becchi, Andrea Ferracani, Alberto Del Bimbo, Liliana Lo Presti, Giuseppe Mazzola, Marco La Cascia, Federico Cunico, Andrea Toaiari, Marco Cristani, Antonio Greco, Alessia Saggese, and Mario Vento. “I-MALL An Effective Framework for Personalized Visits. Improving the Customer Experience in Stores”. In: *Proceedings of the 1st Workshop on Multimedia Computing towards Fashion Recommendation*. MM ’22. ACM, Oct. 2022, pp. 11–19.

[19] Ben Benfold and Ian Reid. “Stable multi-target tracking in real-time surveillance video”. In: *CVPR 2011*. IEEE. IEEE, June 2011, pp. 3457–3464.

[20] Cigdem Beyan, Vasiliki-Maria Katsageorgiou, and Vittorio Murino. “A Sequential Data Analysis Approach to Detect Emergent Leaders in Small Groups”. In: *IEEE Transactions on Multimedia* 21.8 (Aug. 2019), pp. 2107–2116.

[21] Cigdem Beyan, Alessandro Vinciarelli, and Alessio Del Bue. “Co-Located Human–Human Interaction Analysis Using Nonverbal Cues: A Survey”. In: *ACM Computing Surveys* 56.5 (Nov. 2023), pp. 1–41.

[22] Elina Birmingham, Walter F. Bischof, and Alan Kingstone. “Gaze selection in complex social scenes”. In: *Visual Cognition* 16.2–3 (Feb. 2008), pp. 341–355.

[23] Priyaranjan Biswal and Prases K. Mohanty. "Development of quadruped walking robots: A review". In: *Ain Shams Engineering Journal* 12.2 (June 2021), pp. 2017–2031.

[24] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258* (2021).

[25] Ali Borji and Laurent Itti. "State-of-the-Art in Visual Attention Modeling". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (Jan. 2013), pp. 185–207.

[26] Ernesto Brau, Jinyan Guan, Tanya Jeffries, and Kobus Barnard. "Multiple-Gaze Geometry: Inferring Novel 3D Locations from Gazes Observed in Monocular Video". In: *Computer Vision – ECCV 2018*. Springer International Publishing, 2018, pp. 641–659.

[27] Judee K. Burgoon and Adam Kendon. "Conducting Interaction: Patterns of Behavior in Focused Encounters." In: *Contemporary Sociology* 21.2 (Mar. 1992), p. 256.

[28] Yuanhao Cai, Zhicheng Wang, Zhengxiong Luo, Binyi Yin, Angang Du, Haoqian Wang, Xiangyu Zhang, Xinyu Zhou, Erjin Zhou, and Jian Sun. "Learning Delicate Local Representations for Multi-person Pose Estimation". In: *Computer Vision – ECCV 2020*. Springer International Publishing, 2020, pp. 455–472.

[29] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.1 (Jan. 2021), pp. 172–186.

[30] Luigi Capogrosso, Geri Skenderi, Federico Girella, Franco Fummi, and Marco Cristani. "Toward Smart Doors: A Position Paper". In: *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*. Springer. Springer Nature Switzerland, 2023, pp. 660–673.

[31] Luigi Capogrosso, Andrea Toaiari, Andrea Avogaro, Uzair Khan, Aditya Jivoji, Franco Fummi, and Marco Cristani. "SITUATE: Indoor Human Trajectory Prediction through Geometric Features and Self-Supervised Vision Representation". In: *arXiv preprint arXiv:2409.00774* (2024).

[32] Francesca Capozzi, Cigdem Beyan, Antonio Pierro, Atesh Koul, Vittorio Murino, Stefano Livi, Andrew P. Bayliss, Jelena Ristic, and Cristina Becchio. "Tracking the Leader: Gaze Behavior in Group Interactions". In: *iScience* 16 (June 2019), pp. 242–249.

[33] Marisa Carrasco. "Visual attention: The past 25 years". In: *Vision Research* 51.13 (July 2011), pp. 1484–1525.

[34] Kai-Chi Chan, Cheng-Kok Koh, and C. S. George Lee. "Selecting best viewpoint for human-pose estimation". In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. IEEE, May 2014, pp. 4844–4849.

[35] Ching-Hang Chen and Deva Ramanan. "3D Human Pose Estimation = 2D Pose Estimation + Matching". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017.

[36] Yuhao Chen, Yue Luo, Chizhao Yang, Mustafa Ozkan Yerebakan, Shuai Hao, Nicolas Grimaldi, Song Li, Read Hayes, and Boyi Hu. "Human mobile robot interaction in the retail environment". In: *Scientific Data* 9.1 (Nov. 2022), p. 673.

[37] Zhaokang Chen and Bertram E. Shi. "Towards High Performance Low Complexity Calibration in Appearance Based Gaze Estimation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.1 (Jan. 2023), pp. 1174–1188.

[38] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. "HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020, pp. 5385–5394.

[39] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. "A Coarse-to-Fine Adaptive Network for Appearance-Based Gaze Estimation". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.07 (Apr. 2020), pp. 10623–10630.

[40] Yihua Cheng, Feng Lu, and Xucong Zhang. "Appearance-Based Gaze Estimation via Evaluation-Guided Asymmetric Regression". In: *Computer Vision – ECCV 2018*. Springer International Publishing, 2018, pp. 105–121.

[41] Chiho Choi and Behzad Dariush. "Looking to Relations for Future Trajectory Forecast". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019.

[42] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. "Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose". In: *Computer Vision – ECCV 2020*. Springer International Publishing, 2020, pp. 769–787.

[43] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M. Rehg. "Connecting Gaze, Scene, and Attention: Generalized Attention Estimation via Joint Modeling of Gaze and Scene Saliency". In: *Computer Vision – ECCV 2018*. Springer International Publishing, 2018, pp. 397–412.

[44] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M. Rehg. "Detecting Attended Visual Targets in Video". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020, pp. 5395–5405.

[45] Addwiteey Chrungoo, S. S. Manimaran, and Balaraman Ravindran. "Activity Recognition for Natural Human Robot Interaction". In: *Social Robotics*. Springer International Publishing, 2014, pp. 84–94.

[46] Sungho Chun, Sungbum Park, and Ju Yong Chang. "Learnable Human Mesh Triangulation for 3D Human Pose and Shape Estimation". In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Jan. 2023, pp. 2849–2858.

[47] Roberto Cipolla, Yarin Gal, and Alex Kendall. "Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2018, pp. 7482–7491.

[48] Taco Cohen and Max Welling. "Group equivariant convolutional networks". In: *International Conference on Machine Learning*. PMLR. 2016.

[49] Mejdi Dallel, Vincent Havard, David Baudry, and Xavier Savatier. "InHARD - Industrial Human Action Recognition Dataset in the Context of Industrial Collaborative Robotics". In: *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. IEEE, Sept. 2020, pp. 1–6.

[50] Kim M Dalton, Brendon M Nacewicz, Tom Johnstone, Hillary S Schaefer, Morton Ann Gernsbacher, H H Goldsmith, Andrew L Alexander, and Richard J Davidson. "Gaze fixation and the neural circuitry of face processing in autism". In: *Nature Neuroscience* 8.4 (Mar. 2005), pp. 519–526.

[51] Patrick Dendorfer, Sven Elflein, and Laura Leal-Taixe. "MG-GAN: A Multi-Generator Model Preventing Out-of-Distribution Samples in Pedestrian Trajectory Prediction". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2021, pp. 13138–13147.

[52] Furong Deng, Yu Zhou, Sifan Song, Zijian Jiang, Lifu Chen, Jionglong Su, Zhenglong Sun, and Jiaming Zhang. "Say What You Are Looking At: An Attention-Based Interactive System for Autistic Children". In: *Applied Sciences* 11.16 (Aug. 2021), p. 7426.

[53] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. IEEE, June 2009, pp. 248–255.

[54] Philipe A. Dias, Damiano Malafronte, Henry Medeiros, and Francesca Odone. "Gaze Estimation for Assisted Living Environments". In: *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Mar. 2020, pp. 290–299.

[55] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2020.

[56] S. Gareth Edwards, Lisa J. Stephenson, Mario Dalmaso, and Andrew P. Bayliss. “Social orienting in gaze leading: a mechanism for shared attention”. In: *Proceedings of the Royal Society B: Biological Sciences* 282.1812 (Aug. 2015), p. 20151141.

[57] Shirine El Zaatari, Mohamed Marei, Weidong Li, and Zahid Usman. “Cobot programming for collaborative industrial tasks: An overview”. In: *Robotics and Autonomous Systems* 116 (June 2019), pp. 162–180.

[58] N.J. Emery. “The eyes have it: the neuroethology, function and evolution of social gaze”. In: *Neuroscience & biobehavioral reviews* 24.6 (Aug. 2000), pp. 581–604.

[59] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. “Dual Attention Guided Gaze Target Detection in the Wild”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021, pp. 11390–11399.

[60] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. “RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments”. In: *Computer Vision – ECCV 2018*. Springer International Publishing, 2018, pp. 339–357.

[61] Steve Fotios, Jim Uttley, C Cheal, and N Hara. “Using eye-tracking to identify pedestrians’ critical visual tasks, Part 1. Dual task approach”. In: *Lighting research & technology* 47.2 (2015), pp. 133–148.

[62] Tom Foulsham, Esther Walker, and Alan Kingstone. “The where, what and when of gaze allocation in the lab and the natural environment”. In: *Vision Research* 51.17 (Sept. 2011), pp. 1920–1931.

[63] Felix Franzel, Thomas Eiband, and Dongheui Lee. “Detection of Collaboration and Collision Events during Contact Task Execution”. In: *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*. IEEE, July 2021, pp. 376–383.

[64] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. “Se (3)-transformers: 3d roto-translation equivariant attention networks”. In: *Advances in Neural Information Processing Systems* (2020).

[65] JamesH. Fuller. “Eye position and target amplitude effects on human visual saccadic latencies”. In: *Experimental Brain Research* 109.3 (June 1996), pp. 457–466.

[66] Patrizia Gabellini, Mauro D’Aloisio, Matteo Fabiani, and Valerio Placidi. “A Large Scale Trajectory Dataset for Shopper Behaviour Understanding”. In: *New Trends in Image Analysis and Processing – ICIAP 2019*. Springer. Springer International Publishing, 2019, pp. 285–295.

[67] Mercedes Garcia-Salguero, Javier Gonzalez-Jimenez, and Francisco-Angel Moreno. “Human 3D Pose Estimation with a Tilting Camera for Social Mobile Robot Interaction”. In: *Sensors* 19.22 (Nov. 2019), p. 4943.

[68] A. Geiger, P. Lenz, and R. Urtasun. “Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite”. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2012, pp. 3354–3361.

[69] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. "Bottom-Up Human Pose Estimation Via Disentangled Keypoint Regression". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021, pp. 14676–14686.

[70] Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, and Han Hu. "Human Pose as Compositional Tokens". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023, pp. 660–671.

[71] Shreya Ghosh, Abhinav Dhall, Munawar Hayat, Jarrod Knibbe, and Qiang Ji. "Automatic Gaze Analysis: A Survey of Deep Learning Based Approaches". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.1 (Jan. 2024), pp. 61–84.

[72] Francesco Giuliari, Irtiza Hasan, Marco Cristani, and Fabio Galasso. "Transformer Networks for Trajectory Forecasting". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. IEEE, Jan. 2021, pp. 10335–10342.

[73] Robert D. Gordon. "Selective attention during scene perception: Evidence from negative priming". In: *Memory & cognition* 34.7 (Oct. 2006), pp. 1484–1494.

[74] Alex Graves, Greg Wayne, and Ivo Danihelka. "Neural Turing Machines". In: *arXiv preprint arXiv:1410.5401* (2014).

[75] Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. "Stochastic Trajectory Prediction via Motion Indeterminacy Diffusion". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022.

[76] Matthias Guertler, Laura Tomidei, Nathalie Sick, Marc Carmichael, Gavin Paul, Annika Wambsganss, Victor Hernandez Moreno, and Sazzad Hussain. "WHEN IS A ROBOT A COBOT? MOVING BEYOND MANUFACTURING AND ARM-BASED COBOT MANIPULATORS". In: *Proceedings of the Design Society* 3 (June 2023), pp. 3889–3898.

[77] Sheng Guo, Hanjiang Xiong, and Xianwei Zheng. "A Novel Semantic Matching Method for Indoor Trajectory Tracking". In: *ISPRS International Journal of Geo-Information* 6.7 (July 2017), p. 197.

[78] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. "Back to MLP: A Simple Baseline for Human Motion Prediction". In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Jan. 2023, pp. 4798–4808.

[79] Zidong Guo, Zejian Yuan, Chong Zhang, Wanchao Chi, Yonggen Ling, and Shenghao Zhang. "Domain Adaptation Gaze Estimation by Embedding with Prediction Consistency". In: *Computer Vision – ACCV 2020*. Springer International Publishing, 2021, pp. 292–307.

[80] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. "Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2018, pp. 2255–2264.

[81] Anshul Gupta, Samy Tafasca, and Jean-Marc Odobez. "A Modular Multimodal Architecture for Gaze Target Prediction: Application to Privacy-Sensitive Settings". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, June 2022, pp. 5037–5046.

[82] Anshul Gupta, Pierre Vuillecard, Arya Farkhondeh, and Jean-Marc Odobez. "Exploring the Zero-Shot Capabilities of Vision-Language Models for Improving Gaze Following". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, June 2024, pp. 615–624.

[83] Sirin Haddad, Meiqing Wu, He Wei, and Siew Kei Lam. "Situation-aware pedestrian trajectory prediction with spatio-temporal attention model". In: *arXiv preprint arXiv:1902.05437* (2019).

[84] Srijeet Halder, Kereshmeh Afsari, Erin Chiou, Rafael Patrick, and Kaveh Akbari Hamed. "Construction inspection & monitoring with quadruped robots in future human-robot teaming: A preliminary study". In: *Journal of Building Engineering* 65 (Apr. 2023), p. 105814.

[85] Jinni A. Harrigan. "Proxemics, Kinesics, and Gaze". In: *The New Handbook of Methods in Nonverbal Behavior Research*. Oxford University Press, Mar. 2008, pp. 136–198.

[86] Irtiza Hasan, Francesco Setti, Theodore Tsesmelis, Alessio Del Bue, Fabio Galasso, and Marco Cristani. "MX-LSTM: Mixing Tracklets and Vislets to Jointly Forecast Trajectories and Head Poses". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2018, pp. 6067–6076.

[87] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016, pp. 770–778.

[88] Dirk Helbing and Péter Molnár. "Social force model for pedestrian dynamics". In: *Physical Review E* 51.5 (May 1995), pp. 4282–4286.

[89] Young Jin Heo, Dayeon Kim, Woongyong Lee, Hyoungkyun Kim, Jonghoon Park, and Wan Kyun Chung. "Collision Detection for Industrial Collaborative Robots: A Deep Learning Approach". In: *IEEE Robotics and Automation Letters* 4.2 (Apr. 2019), pp. 740–746.

[90] Noriaki Hirose, Dhruv Shah, Ajay Sridhar, and Sergey Levine. "SACSoN: Scalable Autonomous Control for Social Navigation". In: *IEEE Robotics and Automation Letters* 9.1 (Jan. 2024), pp. 49–56.

[91] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851.

[92] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780.

[93] Nora Horanyi, Linfang Zheng, Eunji Chong, Aleš Leonardis, and Hyung Jin Chang. “Where are they looking in the 3D space?” In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, June 2023, pp. 2678–2687.

[94] Wenbo Hu, Changgong Zhang, Fangneng Zhan, Lei Zhang, and Tien-Tsin Wong. “Conditional Directed Graph Convolution for 3D Human Pose Estimation”. In: *Proceedings of the 29th ACM International Conference on Multimedia*. MM ’21. ACM, Oct. 2021.

[95] Yue Hu, Siheng Chen, Ya Zhang, and Xiao Gu. “Collaborative Motion Prediction via Neural Motion Message Passing”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020, pp. 6318–6327.

[96] Zhengxi Hu, Dingye Yang, Shilei Cheng, Lei Zhou, Shichao Wu, and Jingtai Liu. “We Know Where They Are Looking at From the RGB-D Camera: Gaze Following in 3D”. In: *IEEE Transactions on Instrumentation and Measurement* 71 (2022), pp. 1–14.

[97] Zhengxi Hu, Yuxue Yang, Xiaolin Zhai, Dingye Yang, Bohan Zhou, and Jingtai Liu. “GFIE: A Dataset and Baseline for Gaze-Following from 2D to 3D in Indoor Environments”. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023, pp. 8907–8916.

[98] Wenbing Huang, Jiaqi Han, Yu Rong, Tingyang Xu, Fuchun Sun, and Junzhou Huang. “Equivariant Graph Mechanics Networks with Constraints”. In: *International Conference on Learning Representations*. 2021.

[99] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. “STGAT: Modeling Spatial-Temporal Interactions for Human Trajectory Prediction”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019, pp. 6271–6280.

[100] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. “Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7 (July 2014), pp. 1325–1339.

[101] Kalana Ishara, Ivan Lee, and Russell Brinkworth. “Mobile robotic active view planning for physiotherapy and physical exercise guidance”. In: *2015 IEEE 7th International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)*. IEEE. IEEE, July 2015, pp. 130–136.

[102] Tianlei Jin, Qizhi Yu, Shiqiang Zhu, Zheyuan Lin, Jie Ren, Yuanhai Zhou, and Wei Song. “Depth-aware gaze-following via auxiliary networks for robotics”. In: *Engineering Applications of Artificial Intelligence* 113 (Aug. 2022), p. 104924.

[103] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. "Panoptic Studio: A Massively Multiview System for Social Motion Capture". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Dec. 2015, pp. 3334–3342.

[104] Tilke Judd, Krista Ehinger, Fredo Durand, and Antonio Torralba. "Learning to predict where humans look". In: *2009 IEEE 12th International Conference on Computer Vision*. IEEE. IEEE, Sept. 2009, pp. 2106–2113.

[105] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Soren Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. "Socially CompliAnt Navigation Dataset (SCAND): A Large-Scale Dataset of Demonstrations for Social Navigation". In: *IEEE Robotics and Automation Letters* 7.4 (Oct. 2022), pp. 11807–11814.

[106] Sagi Katz, Ayellet Tal, and Ronen Basri. "Direct visibility of point sets". In: *ACM SIGGRAPH 2007 papers*. SIGGRAPH07. San Diego, California: ACM, July 2007, 24–es.

[107] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. "Gaze360: Physically Unconstrained Gaze Estimation in the Wild". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019, pp. 6911–6920.

[108] Diederik P Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[109] Kris M. Kitani, Brian D. Ziebart, James Andrew Bagnell, and Martial Hebert. "Activity Forecasting". In: *Computer Vision – ECCV 2012*. Springer. Springer Berlin Heidelberg, 2012, pp. 201–214.

[110] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. "VIBE: Video Inference for Human Body Pose and Shape Estimation". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020.

[111] Parth Kothari, Sven Kreiss, and Alexandre Alahi. "Human Trajectory Forecasting in Crowds: A Deep Learning Perspective". In: *IEEE Transactions on Intelligent Transportation Systems* 23.7 (July 2022), pp. 7386–7400.

[112] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. "Eye Tracking for Everyone". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016, pp. 2176–2184.

[113] Bernard C. Kress. "Digital optical elements and technologies (EDO19): applications to AR/VR/MR". In: *Digital Optical Technologies 2019*. Ed. by Bernard C. Kress and Peter Schelkens. Vol. 11062. SPIE, July 2019, p. 76.

[114] Murthy L R D and Pradipta Biswas. "Appearance-based Gaze Estimation using Attention and Difference Mechanism". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, June 2021, pp. 3137–3146.

[115] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[116] Andreas M. Lehrmann, Peter V. Gehler, and Sebastian Nowozin. “Efficient Nonlinear Markov Models for Human Motion”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2014, pp. 1314–1321.

[117] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. “Crowds by Example”. In: *Computer Graphics Forum* 26.3 (Sept. 2007), pp. 655–664.

[118] Sijin Li, Weichen Zhang, and Antoni B. Chan. “Maximum-Margin Structured Learning with Deep Networks for 3D Human Pose Estimation”. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Dec. 2015, pp. 2848–2856.

[119] Yin Li, Miao Liu, and James M. Rehg. “In the Eye of the Beholder: Gaze and Actions in First Person Video”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.6 (June 2023), pp. 6731–6747.

[120] Yunhao Li, Wei Shen, Zhongpai Gao, Yucheng Zhu, Guangtao Zhai, and Guodong Guo. “Looking here or there? Gaze Following in 360-Degree Images”. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2021, pp. 3722–3731.

[121] Dongze Lian, Zehao Yu, and Shenghua Gao. “Believe It or Not, We Know What You Are Looking At!” In: *Computer Vision – ACCV 2018*. Springer. Springer International Publishing, 2019, pp. 35–50.

[122] Rongqin Liang, Yuanman Li, Jiantao Zhou, and Xia Li. “STGlow: A Flow-Based Generative Framework With Dual-Graphormer for Pedestrian Trajectory Prediction”. In: *IEEE Transactions on Neural Networks and Learning Systems* 35.11 (Nov. 2024), pp. 16504–16517.

[123] Pin-Ling Liu and Chien-Chi Chang. “Simple method integrating OpenPose and RGB-D camera for identifying 3D body landmark locations in various postures”. In: *International Journal of Industrial Ergonomics* 91 (Sept. 2022), p. 103354.

[124] Zhenguang Liu, Runyang Feng, Haoming Chen, Shuang Wu, Yixing Gao, Yunjun Gao, and Xiang Wang. “Temporal Feature Alignment and Mutual Information Maximization for Video-Based Human Pose Estimation”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 10996–11006.

[125] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. “SMPL: A Skinned Multi-Person Linear Model”. In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. ACM, Aug. 2023, pp. 851–866.

[126] Ilya Loshchilov and Frank Hutter. “Decoupled Weight Decay Regularization”. In: *International Conference on Learning Representations*. 2018.

[127] Feng Lu, Yue Gao, and Xiaowu Chen. “Estimating 3D Gaze Directions Using Unlabeled Eye Images via Synthetic Iris Appearance Fitting”. In: *IEEE Transactions on Multimedia* 18.9 (Sept. 2016), pp. 1772–1782.

[128] Haoyu Ma, Zhe Wang, Yifei Chen, Deying Kong, Liangjian Chen, Xingwei Liu, Xiangyi Yan, Hao Tang, and Xiaohui Xie. “PPT: Token-Pruned Pose Transformer for Monocular and Multi-view Human Pose Estimation”. In: *Computer Vision – ECCV 2022*. Springer Nature Switzerland, 2022, pp. 424–442.

[129] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. “Progressively Generating Better Initial Guesses Towards Next Stages for High-Quality Human Motion Prediction”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 6427–6436.

[130] Praveen Kumar Reddy Maddikunta, Quoc-Viet Pham, Prabadevi B, N Deepa, Kapal Dev, Thippa Reddy Gadekallu, Rukhsana Ruby, and Madhusanka Liyanage. “Industry 5.0: A survey on enabling technologies and potential applications”. In: *Journal of Industrial Information Integration* 26 (Mar. 2022), p. 100257.

[131] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael Black. “AMASS: Archive of Motion Capture As Surface Shapes”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019, pp. 5441–5450.

[132] Barbara Majecka. “Statistical models of pedestrian behaviour in the forum”. MA thesis. School of Informatics, University of Edinburgh, 2009.

[133] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. “It Is Not the Journey But the Destination: Endpoint Conditioned Trajectory Prediction”. In: *Computer Vision – ECCV 2020*. Springer. Springer International Publishing, 2020, pp. 759–776.

[134] Pranav Mantini and Shishir K. Shah. “Human Trajectory Forecasting In Indoor Environments Using Geometric Context”. In: *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*. ICVGIP ’14. ACM, Dec. 2014, pp. 1–8.

[135] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. “History Repeats Itself: Human Motion Prediction via Motion Attention”. In: *Computer Vision – ECCV 2020*. Springer. Springer International Publishing, 2020, pp. 474–489.

[136] Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. “Leapfrog Diffusion Model for Stochastic Trajectory Prediction”. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023, pp. 5517–5526.

[137] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. “Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera”. In: *Computer Vision – ECCV 2018*. Springer International Publishing, 2018, pp. 614–631.

[138] Angel Martiez-Gonzalez, Michael Villamizar, Olivier Canevet, and Jean-Marc Odobez. “Real-time Convolutional Networks for Depth-based Human Pose Estimation”. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. IEEE, Oct. 2018, pp. 41–47.

[139] Roberto Martín-Martín, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. “JRDB: A Dataset and Benchmark of Egocentric Robot Visual Perception of Humans in Built Environments”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.6 (June 2023), pp. 6748–6765.

[140] Soroush Mehraban, Vida Adeli, and Babak Taati. “MotionAGFormer: Enhancing 3D Human Pose Estimation with a Transformer-GCNFormer Network”. In: *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Jan. 2024.

[141] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. “Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision”. In: *2017 International Conference on 3D Vision (3DV)*. IEEE. IEEE, Oct. 2017, pp. 506–516.

[142] David Melcher. “Visual stability”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 366.1564 (Feb. 2011), pp. 468–475.

[143] Wolfgang Merkt, Vladimir Ivan, Yiming Yang, and Sethu Vijayakumar. “Towards Shared Autonomy Applications using Whole-body Control Formulations of Locomanipulation”. In: *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*. IEEE, Aug. 2019, pp. 1206–1211.

[144] Qiaomu Miao, Minh Hoai, and Dimitris Samaras. “Patch-level Gaze Distribution Prediction for Gaze Following”. In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Jan. 2023, pp. 880–889.

[145] Seyed S. Mohammadi, Nuno F. Duarte, Dimitrios Dimou, Yiming Wang, Matteo Taiana, Pietro Morerio, Atabak Dehban, Plinio Moreno, Alexandre Bernardino, Alessio Del Bue, and José Santos-Victor. “3DSGrasp: 3D Shape-Completion for Robotic Grasp”. In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. IEEE, May 2023, pp. 3815–3822.

[146] Jonathan Mumm and Bilge Mutlu. “Human-robot proxemics: physical and psychological distancing in human-robot interaction”. In: *Proceedings of the 6th international conference on Human-robot interaction*. HRI’11. ACM, Mar. 2011, pp. 331–338.

[147] Saeid Nahavandi. “Industry 5.0—A Human-Centric Solution”. In: *Sustainability* 11.16 (Aug. 2019), p. 4371.

[148] Atsushi Nakazawa and Christian Nitschke. “Point of Gaze Estimation through Corneal Surface Reflection in an Active Illumination Environment”. In: *Computer Vision – ECCV 2012*. Springer. Springer Berlin Heidelberg, 2012, pp. 159–172.

[149] Lukáš Neumann and Andrea Vedaldi. "Tiny People Pose". In: *Computer Vision – ACCV 2018*. Springer International Publishing, 2019, pp. 558–574.

[150] Alejandro Newell, Kaiyu Yang, and Jia Deng. "Stacked Hourglass Networks for Human Pose Estimation". In: *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 483–499.

[151] B. Noureddin, P.D. Lawrence, and C.F. Man. "A non-contact device for tracking gaze in a human computer interface". In: *Computer Vision and Image Understanding* 98.1 (Apr. 2005), pp. 52–82.

[152] Jun O Oh, Hyung Jin Chang, and Sang-Il Choi. "Self-Attention with Convolution and Deconvolution for Efficient Eye Gaze Estimation from a Full Face Image". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, June 2022, pp. 4988–4996.

[153] Daniel Parks, Ali Borji, and Laurent Itti. "Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes". In: *Vision Research* 116 (Nov. 2015), pp. 113–126.

[154] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael 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)*. IEEE, June 2019, pp. 7745–7754.

[155] Stefano Pellegrini, Andreas Ess, and Luc Van Gool. "Improving Data Association by Joint Modeling of Pedestrian Trajectories and Groupings". In: *Computer Vision – ECCV 2010*. Springer. Springer Berlin Heidelberg, 2010, pp. 452–465.

[156] Michael I. Posner, Charles R. Snyder, and Brian J. Davidson. "Attention and the detection of signals." In: *Journal of Experimental Psychology: General* 109.2 (June 1980), pp. 160–174.

[157] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. "Habitat 3.0: A co-habitat for humans, avatars and robots". In: *arXiv preprint arXiv:2310.13724* (2023).

[158] Lingteng Qiu, Xuanye Zhang, Yanran Li, Guanbin Li, Xiaojun Wu, Zixiang Xiong, Xiaoguang Han, and Shuguang Cui. "Peeking into Occluded Joints: A Novel Framework for Crowd Pose Estimation". In: *Computer Vision – ECCV 2020*. Springer. Springer International Publishing, 2020, pp. 488–504.

[159] Keith Rayner. "Eye movements in reading and information processing: 20 years of research." In: *Psychological Bulletin* 124.3 (1998), pp. 372–422.

[160] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. "Where are they looking?" In: *Advances in Neural Information Processing Systems* 28 (2015).

[161] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. "Following Gaze in Video". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017, pp. 1444–1452.

[162] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You Only Look Once: Unified, Real-Time Object Detection". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016, pp. 779–788.

[163] Ian Reid, Ben Benfold, Alonso Patron, and Eric Sommerlade. "Understanding Interactions and Guiding Visual Surveillance by Tracking Attention". In: *Computer Vision – ACCV 2010 Workshops*. Springer. Springer Berlin Heidelberg, 2011, pp. 380–389.

[164] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. "Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes". In: *Computer Vision – ECCV 2016*. Springer. Springer International Publishing, 2016, pp. 549–565.

[165] Luca Rossi, Marina Paolanti, Roberto Pierdicca, and Emanuele Frontoni. "Human trajectory prediction and generation using LSTM models and GANs". In: *Pattern Recognition* 120 (Dec. 2021), p. 108136.

[166] Andrey Rudenko, Tomasz P. Kucner, Chittaranjan S. Swaminathan, Ravi T. Chadalavada, Kai O. Arras, and Achim J. Lilienthal. "THÖR: Human-Robot Navigation Data Collection and Accurate Motion Trajectories Dataset". In: *IEEE Robotics and Automation Letters* 5.2 (Apr. 2020), pp. 676–682.

[167] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. "Human motion trajectory prediction: a survey". In: *The International Journal of Robotics Research* 39.8 (June 2020), pp. 895–935.

[168] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. "Trajectron++: Dynamically-Feasible Trajectory Forecasting with Heterogeneous Data". In: *Computer Vision – ECCV 2020*. Springer. Springer International Publishing, 2020, pp. 683–700.

[169] Alessio Sampieri, Guido Maria D'Amely di Melendugno, Andrea Avogaro, Federico Cunico, Francesco Setti, Geri Skenderi, Marco Cristani, and Fabio Galasso. "Pose Forecasting in Industrial Human-Robot Collaboration". In: *Computer Vision – ECCV 2022*. Springer. Springer Nature Switzerland, 2022, pp. 51–69.

[170] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. "Learning to simulate complex physics with graph networks". In: *International Conference on Machine Learning*. PMLR. 2020.

[171] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. "E (n) equivariant graph neural networks". In: *International Conference on Machine Learning*. PMLR. 2021.

[172] Tim Schreiter, Tiago Rodrigues de Almeida, Yufei Zhu, Eduardo Gutierrez Maestro, Lucas Morillo-Mendez, Andrey Rudenko, Tomasz P Kucner, Oscar Martinez Mozos, Martin Magnusson, Luigi Palmieri, et al. "The magni human motion dataset: Accurate, complex, multi-modal, natural, semantically-rich and contextualized". In: *arXiv preprint arXiv:2208.14925* (2022).

[173] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. "Diffusion-Based 3D Human Pose Estimation with Multi-Hypothesis Aggregation". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2023, pp. 14715–14725.

[174] Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Mo Zhou, Zhenxing Niu, and Gang Hua. "SGCN:Sparse Graph Convolution Network for Pedestrian Trajectory Prediction". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021, pp. 8990–8999.

[175] Xiang Shi, You Yang, and Qiong Liu. "I Understand You: Blind 3D Human Attention Inference From the Perspective of Third-Person". In: *IEEE Transactions on Image Processing* 30 (2021), pp. 6212–6225.

[176] Snehesh Shrestha, Yantian Zha, Saketh Banagiri, Ge Gao, Yiannis Aloimonos, and Cornelia Fermuller. "NatSGD: A Dataset with Speech, Gestures, and Demonstrations for Robot Learning in Natural Human-Robot Interaction". In: *arXiv preprint arXiv:2403.02274* (2024).

[177] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. "HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion". In: *International Journal of Computer Vision* 87.1–2 (Aug. 2009), pp. 4–27.

[178] Geri Skenderi, Alessia Bozzini, Luigi Capogrosso, Enrico Carlo Agrillo, Giovanni Perbellini, Franco Fummi, and Marco Cristani. "DOHMO: Embedded Computer Vision in Co-Housing Scenarios". In: *2021 Forum on specification & Design Languages (FDL)*. IEEE. IEEE, Sept. 2021, pp. 01–08.

[179] Geri Skenderi, Luigi Capogrosso, Andrea Toaiari, Matteo Denitto, Franco Fummi, Simone Melzi, and Marco Cristani. "Disentangled Latent Spaces Facilitate Data-Driven Auxiliary Learning". In: *arXiv preprint arXiv:2310.09278* (2023).

[180] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. "Space-Time-Separable Graph Convolutional Network for Pose Forecasting". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2021, pp. 11189–11198.

[181] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. "Deep High-Resolution Representation Learning for Human Pose Estimation". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019, pp. 5686–5696.

[182] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. "Compositional Human Pose Regression". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017, pp. 2621–2630.

[183] Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. "3D Human Pose Estimation with Spatio-Temporal Criss-Cross Attention". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023, pp. 4790–4799.

[184] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. "CSDI: Conditional score-based diffusion models for probabilistic time series imputation". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 24804–24816.

[185] Sibo Tian, Minghui Zheng, and Xiao Liang. "TransFusion: A Practical and Effective Transformer-Based Diffusion Model for 3D Human Motion Prediction". In: *IEEE Robotics and Automation Letters* 9.7 (July 2024), pp. 6232–6239.

[186] Andrea Toaiari, Federico Cunico, Francesco Taioli, Ariel Caputo, Gloria Menegaz, Andrea Giachetti, Giovanni Maria Farinella, and Marco Cristani. "SCENE-pathy: Capturing the Visual Selective Attention of People Towards Scene Elements". In: *Image Analysis and Processing – ICIAP 2023*. Springer. Springer Nature Switzerland, 2023, pp. 352–363.

[187] Andrea Toaiari, Vittorio Murino, Marco Cristani, and Cigdem Beyan. "Upper-Body Pose-based Gaze Estimation for Privacy-Preserving 3D Gaze Target Detection". In: *arXiv preprint arXiv:2409.17886* (2024).

[188] Francesco Tonini, Cigdem Beyan, and Elisa Ricci. "Multimodal Across Domains Gaze Target Detection". In: *Proceedings of the 2022 International Conference on Multimodal Interaction*. ICMI '22. ACM, Nov. 2022, pp. 420–431.

[189] Francesco Tonini, Nicola Dall'Asen, Cigdem Beyan, and Elisa Ricci. "Object-aware Gaze Target Detection". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2023, pp. 21803–21812.

[190] Alexander Toshev and Christian Szegedy. "DeepPose: Human Pose Estimation via Deep Neural Networks". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2014, pp. 1653–1660.

[191] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. "Training data-efficient image transformers & distillation through attention". In: *International Conference on Machine Learning*. PMLR. 2021.

[192] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. "Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors". In: *Procedings of the British Machine Vision Conference 2017*. BMVC 2017. British Machine Vision Association, 2017, pp. 1–13.

[193] Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. "End-to-End Human-Gaze-Target Detection with Transformers". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. IEEE, June 2022, pp. 2192–2200.

[194] R. Valenti, N. Sebe, and T. Gevers. "Combining Head Pose and Eye Location Information for Gaze Estimation". In: *IEEE Transactions on Image Processing* 21.2 (Feb. 2012), pp. 802–815.

[195] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017.

[196] Edward Vendrow, Duy Tho Le, Jianfei Cai, and Hamid Rezatofighi. "JRDB-Pose: A Large-Scale Dataset for Multi-Person Pose Estimation and Tracking". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023, pp. 4811–4820.

[197] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. "Social signal processing: Survey of an emerging domain". In: *Image and Vision Computing* 27.12 (Nov. 2009), pp. 1743–1759.

[198] Binglu Wang, Tao Hu, Baoshan Li, Xiaojuan Chen, and Zhijie Zhang. "GaTector: A Unified Framework for Gaze Object Prediction". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 19566–19575.

[199] J.M. Wang, D.J. Fleet, and A. Hertzmann. "Gaussian Process Dynamical Models for Human Motion". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.2 (Feb. 2008), pp. 283–298.

[200] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. "Deep 3D human pose estimation: A review". In: *Computer Vision and Image Understanding* 210 (Sept. 2021), p. 103225.

[201] Peng Wang, Jing Yang, and Jianpei Zhang. "Indoor Trajectory Prediction for Shopping Mall via Sequential Similarity". In: *Information* 13.3 (Mar. 2022), p. 158.

[202] Peng Wang, Jing Yang, and Jianpei Zhang. "Location Prediction for Indoor Spaces based on Trajectory Similarity". In: *2021 4th International Conference on Data Science and Information Technology*. DSIT 2021. ACM, July 2021, pp. 402–407.

[203] Ping Wei, Yang Liu, Tianmin Shu, Nanning Zheng, and Song-Chun Zhu. "Where and Why are They Looking? Jointly Inferring Human Attention and Intentions in Complex Tasks". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2018, pp. 6801–6809.

[204] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. "Convolutional Pose Machines". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016.

[205] Conghao Wong, Beihao Xia, Ziqian Zou, Yulong Wang, and Xinge You. "SocialCircle: Learning the Angle-based Social Interaction Representation for Pedestrian Trajectory Prediction". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2024, pp. 19005–19015.

[206] Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. "GroupNet: Multiscale Hypergraph Neural Networks for Trajectory Prediction with Relational Reasoning". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 6488–6497.

[207] Chenxin Xu, Weibo Mao, Wenjun Zhang, and Siheng Chen. "Remember Intentions: Retrospective-Memory-based Trajectory Prediction". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 6478–6487.

[208] Chenxin Xu, Robby T. Tan, Yuhong Tan, Siheng Chen, Yu Guang Wang, Xinchao Wang, and Yanfeng Wang. "EqMotion: Equivariant Multi-Agent Motion Prediction with Invariant Interaction Reasoning". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023, pp. 1410–1420.

[209] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. "Deep Kinematics Analysis for Monocular 3D Human Pose Estimation". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020, pp. 896–905.

[210] Pei Xu, Jean-Bernard Hayet, and Ioannis Karamouzas. "SocialVAE: Human Trajectory Prediction Using Timewise Latents". In: *Computer Vision – ECCV 2022*. Springer. Springer Nature Switzerland, 2022, pp. 511–528.

[211] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. "ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation". In: *Advances in Neural Information Processing Systems*. 2022.

[212] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. "MT-VAE: Learning Motion Transformations to Generate Multimodal Human Dynamics". In: *Computer Vision – ECCV 2018*. Springer International Publishing, 2018, pp. 276–293.

[213] Zhi Yan, Li Sun, Tom Duckctr, and Nicola Bellotto. "Multisensor Online Transfer Learning for 3D LiDAR-Based Human Detection with a Mobile Robot". In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Oct. 2018, pp. 7635–7640.

[214] Yi Yang and Deva Ramanan. "Articulated Human Detection with Flexible Mixtures of Parts". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.12 (Dec. 2013), pp. 2878–2890.

[215] Yiding Yang, Zunlei Feng, Mingli Song, and Xinchao Wang. "Factorizable graph convolutional networks". In: *Advances in Neural Information Processing Systems* (2020).

[216] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. "Spatio-Temporal Graph Transformer Networks for Pedestrian Trajectory Prediction". In: *Computer Vision – ECCV 2020*. Springer. Springer International Publishing, 2020, pp. 507–523.

[217] Ye Yuan and Kris Kitani. "DLow: Diversifying Latent Flows for Diverse Human Motion Prediction". In: *Computer Vision – ECCV 2020*. Springer. Springer International Publishing, 2020, pp. 346–364.

[218] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris Kitani. "AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2021, pp. 9793–9803.

[219] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. "Hrformer: High-resolution vision transformer for dense predict". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 7281–7293.

[220] Andrei Zanfir, Mihai Zanfir, Alex Gorban, Jingwei Ji, Yin Zhou, Dragomir Anguelov, and Cristian Sminchisescu. "HUM3DIL: Semi-supervised Multi-modal 3D Human Pose Estimation for Autonomous Driving". In: *Proceedings of The 6th Conference on Robot Learning*. Ed. by Karen Liu, Dana Kulic, and Jeff Ichnowski. Vol. 205. Proceedings of Machine Learning Research. PMLR, Dec. 2023, pp. 1114–1124.

[221] Arthur Zhang, Chaitanya Eranki, Christina Zhang, Ji-Hwan Park, Raymond Hong, Pranav Kalyani, Lochana Kalyanaraman, Arsh Gamare, Arnav Bagad, Maria Esteva, et al. "Towards Robust Robot 3D Perception in Urban Environments: The UT Campus Object Dataset". In: *arXiv preprint arXiv:2309.13549* (2023).

[222] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. "MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 13222–13232.

[223] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. "SR-LSTM: State Refinement for LSTM Towards Pedestrian Trajectory Prediction". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019, pp. 12077–12086.

[224] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. "S^3FD: Single Shot Scale-Invariant Face Detector". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017, pp. 192–201.

[225] Xiaoxiong Zhang, Adarsh Ghimire, Sajid Javed, Jorge Dias, and Naoufel Werghi. "Robot-Person Tracking in Uniform Appearance Scenarios: A New Dataset and Challenges". In: *IEEE Transactions on Human-Machine Systems* 53.3 (June 2023), pp. 549–559.

[226] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. "ETH-XGaze: A Large Scale Dataset for Gaze Estimation Under Extreme Head Pose and Gaze Variation". In: *Computer Vision – ECCV 2020*. Springer International Publishing, 2020, pp. 365–381.

[227] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. "Evaluation of Appearance-Based Methods and Implications for Gaze-Based Applications". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. ACM, May 2019, pp. 1–13.

[228] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. "Appearance-based gaze estimation in the wild". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015, pp. 4511–4520.

[229] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. "It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, July 2017, pp. 51–60.

[230] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. "MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.1 (Jan. 2019), pp. 162–175.

[231] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. "ByteTrack: Multi-object Tracking by Associating Every Detection Box". In: *Computer Vision – ECCV 2022*. Springer Nature Switzerland, 2022, pp. 1–21.

[232] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. "PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023, pp. 8877–8886.

[233] Ce Zheng, Matias Mendieta, Taojiannan Yang, Guo-Jun Qi, and Chen Chen. "FeatER: An Efficient Network for Human Reconstruction via Feature Map-Based TransformER". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023, pp. 13945–13954.

[234] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. "Deep Learning-based Human Pose Estimation: A Survey". In: *ACM Computing Surveys* 56.1 (Aug. 2023), pp. 1–37.

[235] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. "3D Human Pose Estimation with Spatial and Temporal Transformers". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2021, pp. 11636–11645.

[236] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. "MotionBERT: A Unified Perspective on Learning Human Motion Representations". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2023, pp. 15039–15053.

[237] Yue Zhu, Nermin Samet, and David Picard. “H3WB: Human3.6M 3D Whole-Body Dataset and Benchmark”. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2023, pp. 20109–20120.

[238] Christian Zimmermann, Tim Welschehold, Christian Dornhege, Wolfram Burgard, and Thomas Brox. “3D Human Pose Estimation in RGBD Images for Robotic Task Learning”. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. IEEE, May 2018, pp. 1986–1992.