OCCLUSION-AWARE BENCHMARKING IN 3D HUMAN POSE AND SHAPE ESTIMATION

by

Emre Girgin

B.S., Computer Engineering, Boğaziçi University, 2021

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2024

ACKNOWLEDGEMENTS

I would like to dedicate this thesis to my loving wife, Tuba. Things would not be possible without her.

I would like to thank my advisor Prof. Lale Akarun and Assist. Prof. Berk Gökberk to their guidance.

ABSTRACT

OCCLUSION-AWARE BENCHMARKING IN 3D HUMAN POSE AND SHAPE ESTIMATION

3D human pose and shape reconstruction is a widely studied area in computer vision. In addition to non-rigid features and highly articulated joints, another challenge in this area is occlusion, which is common in nature. Although some methods explicitly try to handle occlusion cases, the benchmark against which they are evaluated is vague.

The typical approach in the literature is to report the performance of the method on an occlusion-oriented subset. However, to form such a subset, it is necessary to quantify the occlusion in the samples. The existing approach uses the keypoints and bounding boxes to quantify and rank samples based on occlusion. However, it fails in several cases and tends to produce false positives.

This study proposes the Occlusion Index, a novel index to quantify occlusion in images with high accuracy. The instance mask-based approach not only successfully quantifies occlusion, but also discriminates between occluders and occluders. It also reports the self-occlusion of a person, which is an unavoidable phenomenon in single-view reconstruction.

The experiments show the superiority of the Occlusion Index by forming more challenging subsets, causing state-of-the-art occlusion-robust methods to fail more often. Also, some of the most occluded samples in the popular 3D human pose and shape estimation datasets are included.

ÖZET

3B İNSAN POZU VE ŞEKLİ TAHMİNİNDE ÖRTME DUYARLI KIYASLAMA

3B insan pozu ve şekli elde edilmesi, bilgisayarlı görüde yaygın olarak çalışılan bir alandır. Rijit olmayan özellikler ve yüksek mafsallı eklemlere ek olarak, bu alandaki bir diğer zorluk da doğada yaygın olarak görülen örtmedir. Bazı yöntemler örtme durumlarını ayrıca ele almaya çalışsa da, değerlendirildikleri ölçüt yeterince incelenmemiştir.

Literatürdeki tipik yaklaşım, yöntemin performansını örtme odaklı bir alt küme üzerinde raporlamaktır. Ancak böyle bir alt küme oluşturmak için örneklerdeki örtmeyi ölçmek gerekir. Mevcut yaklaşım örtmeye dayalı örnekleri ölçmek ve sıralamak için anahtar noktaları ve sınırlayıcı kutuları kullanmaktadır. Bununla birlikte, açıkça başarısız olduğu durumlar vardır ve yanlış pozitifler üretme eğilimindedir.

Bu çalışma, görüntülerdeki örtmeyi yüksek doğrulukla ölçmek için yeni bir endeks olan Örtme Endeksini önermektedir. Bireysel maske tabanlı yaklaşım sadece örtmeyi başarılı bir şekilde ölçmekle kalmaz, aynı zamanda örtülenler ile örtenler arasında ayrım yapma kabiliyetine sahiptir. Ayrıca, tek görüntüden yeniden yapılandırmada kaçınılmaz bir durum olan kendi kendine örtmeyi de ölçebilmektedir.

Deneylerde Örtme Endeksi daha zorlu alt kümeler oluşturarak üstün olduğunu göstermiş ve en gelişmiş örtmeye dayanıklı yöntemlerin dahi başarısında düşüşe neden olmuştur. Ayrıca, popüler 3B insan pozu ve şekli veri kümelerindeki örtme değeri en yüksek örneklerin bazıları verilmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

c^j	Pixel color of the jth body segment
d^j	Function producing occlusion weight of joint j
d^{jq}	Indicator function for joint j in group q
I_i^o	Occlusion mask image of ith person
I_i^s	Silhouette mask image of ith person
k_i	Set of occluded joints of person i
k_i^q	Set of occluded joints of person i within group q
k_i^j	jth joint of the ith person
K_i	Set of all joints belonging to ith person
K_i^q	Set of all joints belonging to ith person within group q
m	Number of joints
m^q	Number of joints in group q
M_i	Instance mask of ith person
N_i^a	The number of joint belonging to ith person
N_i^b	The number of joint not belonging to ith person
n	Number of people in the frame
o_i^j	The occlusion score of jth body segment of person i
o_i^q	The occlusion score of group q for person i
O_i	Set of occluded pixels of ith person
S_i	Set of silhouette pixels of ith person
t^j	The truncation score of jth body segment
v^{jq}	Axis angle rotation prediction of jth joint's qth dimension
w^{jq}	Axis angle rotation groundtruth of jth joint's qth dimension

LIST OF ACRONYMS/ABBREVIATIONS

2D	Two Dimensional
3D	Three Dimensional
3DOH50K	3D Occluded Human 50K
3DPW	3D Poses-in-the-Wild
AGORA	Avatars in Geography Optimized for Regression Analysis
AMASS	Archive of Motion Capture as Surface Shapes
BMP	Body Meshes as Points
BEDLAM	Bodies Exhibiting Detailed Lifelike Animated Motion
BEV	Bird Eye View
CI	Crowd Index
CNN	Convolutional Neural Networks
COCO	Common Objects In Context
CoNorm	Context Normalization
DMHS	Deep Multitask Automatic 2D and 3D Human Sensing
EFT	Exemplar Fine Tuning
GCNN	Graph Convolutional Neural Networks
GLAMR	Global Occlusion-Aware Human Mesh Recovery
GTA	Grand Theft Auto
GRU	Gated Recurrent Unit
GraphCMR	Graph Convolutional Mesh Regression
HMR	Human Mesh Recovery
HPS	Human Pose and Shape
ImpHMR	Implicit 3D Human Mesh Recovery
IMU	Inertial Measurement Units
IoU	Intersection over Union
LIDAR	Light Detection and Ranging
LSP	Leeds Sports Pose - Extended
MPII	Max Planck Institute II

MPI-INF-3DHP	Max Planck Institute for Informatics 3D Human Pose
MPJPE	Mean Per Joint Position Error
OCHMR	Occluded Human Mesh Reconstruction
OCHuman	Occluded Human
OI	Occlusion Index
PA	Procrustes Alignment
PA-MPJPE	Procrustes Aligned Mean Per Joint Position Error
PARE	Part Attention Regressor for 3D Human Body Estimation
PASCAL VOC	PASCAL Visual Object Classes
PCA	Principal Component Analysis
RCNN	Recurrent Convolutional Neural Networks
RGB	Red-Green-Blue
ROI	Regional Occlusion Index
ROMP	Regression of Multiple 3D People
SDK	Software Development Kit
SMIL	Skinned Multi-Infant Linear Model
SMPL	Skinned Multi-Person Linear Model
SURREAL	Synthetic Humans for Real Tasks
SPIN	SMPL Optimization in the Loop
SSP-3D	Sports Shape and Pose 3D
ToF	Time-of-Flight
UP-3D	United People 3D
VIBE	Video Inference for Human Body Pose and Shape Estimation

1. INTRODUCTION

3D modeling of objects from images and videos has attracted the attention of computer vision researchers [1–5]. As a result, modeling of rigid objects such as buildings, cityscapes, and household objects, as well as non-rigid objects such as animals and human bodies, has been widely studied [1], [4], [6,7]. The 3D format allows researchers to study the visual appearance of the object and carries much more information than 2D structures. While surface features are prominent, other physical properties such as size rigidness, and stiffness can also be observed.

One debate about 3D reconstruction is the nature of the information source. Since a 3D structure in nature cannot be directly transferred to computers, the type of data collected from nature is also important. While specialized hardware can scan an object in 3D [1], [8], they are financially and computationally expensive and require a carefully designed studio leading to a synthetic environment, making it difficult for them to become the standard. On the other hand, more common and cheaper ToF hardware such as LIDAR provides some 3D elements [9], but this data must be processed to obtain a solid surface topology.

Therefore, the most prominent candidate as a source of information gathered from nature is the 2D images due to their abundance and relatively low cost. While using 2D images for 3D information, the trivial approach is to take multiple snapshots of the objects from different viewpoints to cover all sides [10,11] or to record video-like temporal data around them [7], [12]. The methods using multi-view geometry [13] and structure from motion [7] algorithms are able to reconstruct the object, but this setup may not be available for all cases. Therefore, the remaining option is to reconstruct from a single image, which has gained more attention in recent studies [3], [14–19]. Although 3D reconstruction from a single RGB image is challenging by definition due to invisible sides and lack of depth sense leading to loss of information, recent studies have reported promising results.

Human surface model restoration has a unique place among the options listed above, known as 3D human pose and shape (HPS) estimation. The 3D characteristics of the human body vary greatly depending on age, weight, and the action being performed by the agent. Another major challenge for a human surface model is the non-rigid nature of the body. The human body is highly articulated and the soft surface tissues are deformable. Therefore, maintaining surface consistency under extreme poses is a complex problem [20]. Pose-invariant recovery of the surface model is one of the key goals in this area [21]. Figure 1.1 shows an example image(a) and the output of the 3D HPS estimation(b).

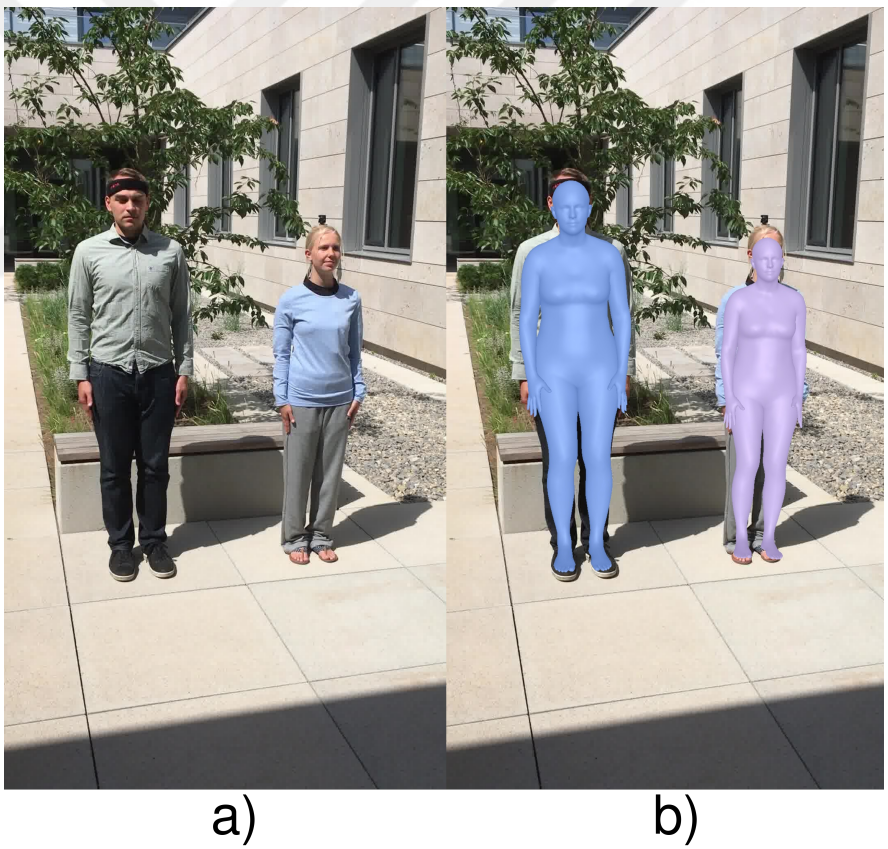

Figure 1.1. An example 3D HPS input(a) and output(b).

In addition to articulation, occlusion is a very common challenge in the wild [17], [21–23]. Human bodies can be occluded by either an object [5] or another person [22], which causes information about the person's 3D pose and orientation to be lost or hidden, leading to inadequate performance for reconstruction methods. This problem

is most often unavoidable when multiple people interact with each other. Figure 1.2 shows an example of the failure of 3D HPS methods due to occlusion. From left to right, Figure 1.2 demonstrates the input image (a), the 3D HPS estimation model's output (b), and the model's output as a bird's eye view (c), respectively. While the most common result of an occlusion failure is not finding the person, in this case the model found an additional 3D model. Also, none of the 3D models that are supposed to be the occluded person fit the person correctly. The closer one (turquoise) is smaller than it should be, the farther one (dark blue) is placed too far away while it should be in touch with the occluding person (purple).

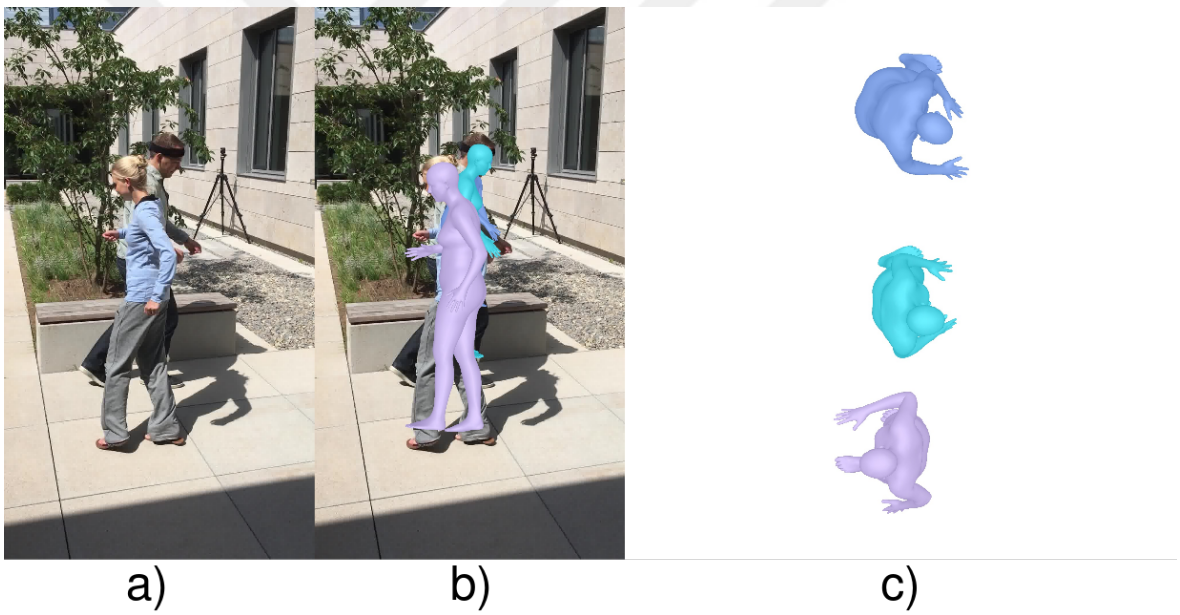

Figure 1.2. An example failure due to occlusion with input(a), output(b), and bird-eye-view of output(c).

Due to the ill-posed nature of the problem and the difficulty in obtaining a 3D human mesh from a single image, recent studies have predominantly utilized deep learning methods [22], [24, 25]. While these state-of-the-art techniques have shown remarkable success in recovering human meshes from a single image, they have often experienced performance degradation when faced with severe occlusion [4], [6], [18]. New studies dealing with object or person occlusion have emerged to overcome the limitations of previous work [5], [22], and they have attracted considerable attention.

Various techniques have been proposed to handle occlusion explicitly in studies. These techniques can be categorized into several categories. One of the newest and popular methods is to simultaneously localize humans in the image as body center heatmaps [15], [17], [22, 23] and perform 3D mesh regression. Earlier methods relied on 3D human pose, which was also calculated concurrently in the neural network architecture [26, 27]. Another common approach to image localization is to divide the image into a grid and analyze each cell based on the probability of a human being present [17], [23]. It is important to note that while this method is a form of localization, some techniques combine grids and body center heatmaps [17], [22,23]. Although spatial features are typically used, some studies incorporate context and expected pose using an attention mechanism [28].

However, it is important to note that the datasets used for training and model evaluation are just as crucial as the robustness of deep learning methods. In the last decade, many human network datasets have been published [29–32]. However, due to the difficulty in acquiring 3D annotations, the datasets are categorized by the method used. The most common approach is to use a third-party method that estimates the 3D human mesh [25]. This technique without markers is susceptible to incorrect annotations due to the limitations of the third-party method. Another option is to use humans with reflector markers that provide additional location information [30], [33,34]. Although these datasets provide more accurate annotations and are more prestigious than markerless ones, they require additional hardware, including both reflectors and receivers. These datasets differ in the hardware they use. While some datasets have been collected in the wild using small hardware [30], [34], most marker-based datasets require more complex and significant instruments [1], [8], limiting them to artificial environments such as laboratories. In the second iteration, a human annotator can correct fine details in both markerless and marker-based datasets. In addition to datasets collected through these approaches, datasets generated through simulation are also popular [29], [32], [35]. Some studies have collected data by placing pre-scanned 3D human models in machine-generated scenes because their 3D structure is already known. Synthetic datasets provide more accurate annotations, but they suffer from

a distribution gap between synthetic and real-world scenarios. Finally, algorithms for estimating 3D human pose and shape utilize auxiliary datasets that may not include 3D surface annotations, but instead contain other annotations such as 2D/3D poses or silhouettes [33], [36–38].

The current approach involves creating a subset of benchmark datasets to compare model robustness against occlusion [3], [17]. Therefore, the selection criteria for sampled frames in the dataset are critical. For instance, collecting a subset where the samples are not as challenging as they should be will adversely affect the evaluation of occlusion robustness performance. Such errors can lead to the illusion of having an occlusion robust 3D HPS estimator. The current practice of using bounding boxes and 2D keypoints for comparison leads to an inaccurate ranking between frames regarding the amount of occlusion, resulting in an unfair comparison.

This study aims to classify and rank samples in the dataset based on the amount and type of occlusion in RGB images. The objective is to investigate the effect of these challenges and propose techniques to improve the performance of existing and future work on the dataset.

This study proposes Occlusion Index to assess occlusion in datasets [39]. The existing index for determining the amount of occlusion in a frame has several drawbacks, such as ignoring object occlusion, inaccurately measuring the overlap between the limbs of different agents, and losing the hierarchy between occluders and occluded persons. Occlusion Index (OI) [39] considers the depth order between agents and provides a more sensitive measure due to its segmentation mask-based approach. The experiments support the superiority of our index by showing that the samples selected by it challenge the state-of-the-art methods more than those selected by the existing metric.

We extended the Occlusion Index to measure the model's sensitivity to the occluded body part. The dataset was divided into subsets based on the occlusion location, and the samples in each subset were ranked based on the amount of occlusion. Since

each body part carries different amounts of information about the person's pose due to the different articulation capabilities of the body segments, each model may be more sensitive to the occlusion of specific subsets. It has been demonstrated that the performance of state-of-the-art methods is influenced by various body segments, as anticipated.

To address the limitations of assigning binary weights for keypoint visibility, we directly utilized instance masks instead of 2D keypoints. We achieved this by rendering each 3D human model with different colors, resulting in a pseudo-instance mask for each model in the frame. Additionally, we obtained a mask per body segment by rendering each segment separately in the same manner. The relationship between the occlusion ratio of a body segment and the error generated for that model indicates the importance of the visibility of that segment. We applied this approach to the entire dataset and determined the importance of each segment's visibility. We also demonstrated the transferability and similarity of these weights across multiple models and datasets.

Fourth, we investigated the effect of self-occlusion, since single-view reconstruction always suffers from occlusion. We examined the correlation between the average error and the body orientation. Then, we examined the visibility of each body segment separately for self-occlusion cases.

Finally, we propose a new comparison metric that evaluates the performance of the methods for 3D human mesh reconstruction tasks. Regressing the pose of an invisible joint is a problem with an unclear solution. While a model may provide a reasonable pose for that segment, the annotation for that part is only an approximation made by an annotator. Therefore, instead of attempting to predict the annotator's thoughts on this joint, we propose a modified metric that evaluates the accuracy of the visible parts. This modification enables researchers to evaluate their model's performance solely on the visible parts, rather than on unknown segments, to determine if the model's performance on these visible regions is affected by occlusion elsewhere.

2. LITERATURE REVIEW

Since the emergence of human pose and shape (HPS) estimation as one of the main research areas, many efforts have been made in this field. This chapter surveys the extensive body of literature in this domain, exploring the diverse methodologies and datasets that have contributed to our understanding of human pose and shape estimation. Furthermore, this text critically examines the new approaches introduced to increase the robustness of these methods to occlusion, a crucial challenge in real-world scenarios where individuals are frequently partially concealed from view. This chapter serves as a foundation for comprehending the state of the art and identifying opportunities for further research and development in the field of occlusion handling, by exploring the evolution of techniques, benchmark datasets, and the varying degrees of success.

2.1. A Skinned Multi-Person Linear Model: SMPL

A key contribution that has influenced recent research directions is the introduction of a parametric human body model called the Skinned Multi-Person Linear Model (SMPL) [40]. This learned model incorporates pose-dependent shape variations and regressed joint information. By using Principal Component Analysis (PCA) [41] to reduce the dimensionality of the parameter space, this model can efficiently reconstruct a wide range of human body meshes with only 72 parameters for pose, 10 parameters for shape, and three parameters for translation. SMPL's versatility and impact have been demonstrated through its various adaptations in diverse domains. The joint list of SMPL is the following order: *Pelvis*, *Left Hip*, *Right Hip*, *Spine1*, *Left Knee*, *Right Knee*, *Spine2*, *Left Ankle*, *Right Ankle*, *Spine3*, *Left Foot*, *Right Foot*, *Neck*, *Left Collar*, *Right Collar*, *Head*, *Left Shoulder*, *Right Shoulder*, *Left Elbow*, *Right Elbow*, *Left Wrist*, *Right Wrist*, *Left Hand*, and *Right Hand*. The SMPL pose parameters are represented as the relative rotation of joints in axis-angle format. The corresponding joints are visualized in Figure 2.1.

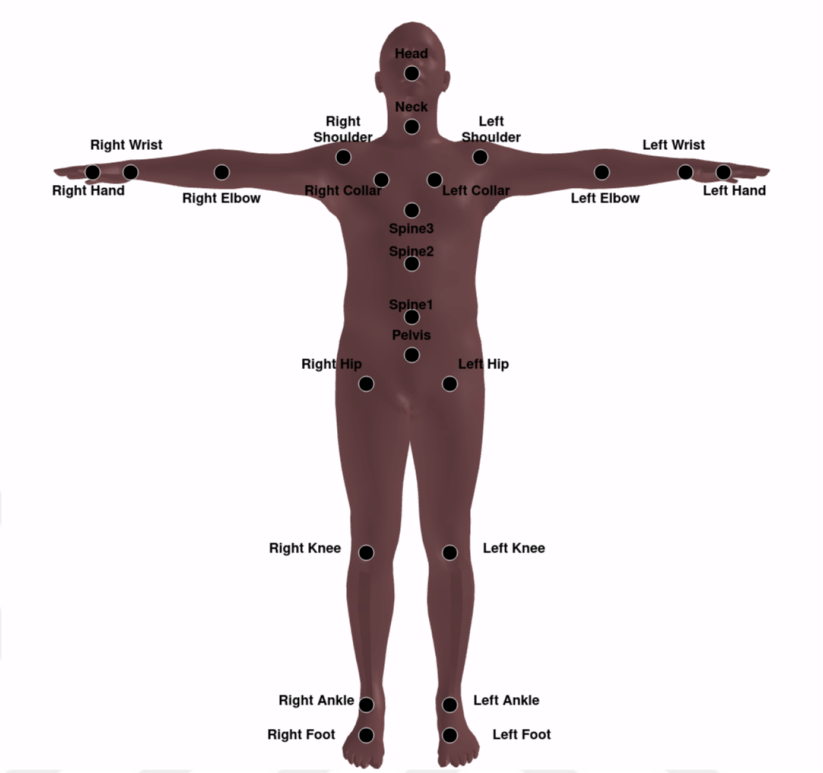

Figure 2.1. Corresponding joints of SMPL pose parameters.

2.2. Datasets

As deep learning techniques become the dominant approach for human pose and shape estimation, there is a growing need for comprehensive and meticulously annotated datasets. However, obtaining precise 3D models of human agents remains a challenge. Many studies use third-party algorithms to fit a 3D human model to 2D features. However, some endeavors require specialized hardware, which can result in the collection of data that represents unnatural scenarios. Most datasets rely on 3D annotations provided by human annotators, but some studies use marker-based tracking methods to infer the poses and shapes of agents by fitting markers placed on them. The variety of data collection methodologies highlights the changing landscape of dataset creation in this field.

Datasets that contain 3D human meshes or parameters to derive them as ground-truth exhibit variations in how they obtain the 3D mesh. Approaches that involve

placing motion-capture (MoCap) markers or sensors directly on the agent provide a degree of supervision, making them a natural focal point for research. However, they still require additional steps for marker fitting, which could introduce inaccuracies in the annotations. Methods that rely solely on physical markers may use existing techniques to fit the input image and obtain confirmation from human annotators.

Another common method is to use simulations to create synthetic datasets, which have the benefit of providing readily available 3D mesh information. However, using synthetic environments presents challenges, such as a distribution shift and domain gap when applying these datasets in real-world scenarios. Additionally, when SMPL parameters are necessary, further fitting steps are usually needed to extract them from the mesh. This is often the case. Lastly, datasets generated through simulations depend on third-party algorithms for labeling instances. However, they benefit from direct supervision of the 3D mesh provided by the simulation software, which represents the highest level of annotation quality.

Furthermore, the HSP task involves both pose estimation and shape fitting. To enhance the robustness of these methods, datasets that contain pose or silhouette ground-truth information without mesh surface are also utilized.

2.2.1. Marker-based Datasets

Marker-based data collection introduces supplementary supervision alongside the spatial 2D features. However, datasets that rely on markers are highly dependent on the hardware used for data acquisition. For example, the use of large reflector markers, commonly employed in the film industry, can disrupt the natural appearance of the agent in the captured images and often requires data collection within controlled environments, such as laboratories. On the other hand, smaller sensors such as Inertial Measurement Units (IMUs) can be discreetly concealed by clothing, making them suitable for in-the-wild scenarios. However, this discretion comes at the cost of weaker signal acquisition. The fitting process, which relies on both the 2D image and sensor

data, may need to be revised. This problem is generally caused by the noise present in nature that significantly increases when there is multi-person interaction, like the one in the multi-person 3D human pose and shape estimation.

Collecting 3D annotations from nature and human surface shape is a challenging task. One approach to annotating the data is to adopt a third-party method. However, this approach has limitations that can affect the accuracy of the annotation.

Novel methods can address this limitation and improve the quality of the annotation by replicating the reflective marker idea in real-world scenarios. One popular approach is to attach Inertial Measurement Units (IMUs) to the actors' bodies in the scene.However, associating these IMU signals with the actors in the frames requires additional processing.

The 3D Poses-in-the-Wild (3DPW) [30] dataset is one of the most widely used resources in the field. True to its name, this dataset consists of in-the-wild video sequences captured with handheld smartphones featuring IMU-equipped actors engaged in various activities. An example frame from the 3DPW dataset is shown in Figure 2.2.The person on the right has a bandage on their head, and the person on the left has a bandage on their neck, under which the IMU sensors are placed.

To align each 2D skeleton in the image space with its corresponding IMU, 3DPW utilizes a graph-based optimization that considers both the image and IMU sequences. This optimization jointly optimizes SMPL parameters, heading drift, and camera parameters while also ensuring the long-term association between IMU data and the video sequence. To validate the accuracy of the framework, the Total Capture [34] benchmark, which provides IMU and human pose data, is utilized. They provide a dataset consisting of 60 distinct sequences and more than 51,000 frames annotated with 2D/3D pose, SMPL parameters, and camera poses. This dataset is a comprehensive and invaluable resource for research in human pose and shape estimation and it is widely utilized in many 3D computer vision problems related to humans.

Figure 2.2. An example image of 3DPW dataset.

Other marker-based datasets commonly used for 3D human pose estimation include Human3.6M [33], HumanEva I/II [36], and Total Capture [34]. Human3.6M utilizes reflective markers in a lab environment to provide 3D human pose data. The dataset includes 3.6 million frames from four different viewpoints and 15 actions. However, it does not directly provide SMPL parameters. These parameters are generated using MoSh [42] and SMPLify-X [43]. HumanEva I and II offer multiview videos of humans with reflective markers, but do not include SMPL parameters. The Total Capture Dataset provides signals from IMUs placed on humans that are synchronized with the video, but does not include SMPL parameters either. AMASS [44] (Archive of Motion Capture as Surface Shapes) is a popular database containing many marker-based datasets. Please refer to Table 2.1 for a list of the mentioned marker-based datasets.

Note that only the datasets that have often been utilized for 3D human pose and shape estimation are included.

Table 2.1. Marker-based datasets comparison table.

Name	# Frames	SMPL	In-the-wild	Marker Type
3DPW	51K	Present	Yes	IMU
Human3.6M	3.6M	Third-Party	No	Reflective
Total Capture	1.9M	Absent	No	IMU
HumanEva I/II	80K	Absent	No	Reflective

2.2.2. Markerless Datasets

Markerless methods require the use of 3D algorithms for Human Pose and Shape (HPS) to extract and analyze pose and shape information from 2D images. These algorithms often build upon previous works. However, relying solely on existing methodologies may limit the quality of annotations. Additionally, annotations may be less reliable when dealing with occlusion scenarios. To address these challenges, some datasets use human annotators to meticulously refine annotation details, thereby improving the accuracy and fidelity of the data.

The Occluded-Human (OCHuman) [31] dataset is a benchmark specifically curated to address complex multi-person occlusion scenarios. It comprises images from real-world settings where human poses are often highly articulated, posing a significant challenge. The dataset includes 4,731 images, with instances selected based on a minimum person-to-person MaxIoU score of 0.5 to qualify as instances of heavy occlusion. On average, the dataset maintains a MaxIoU of 0.67 across all instances. Annotation details for each image include bounding boxes, segmentation masks, and 17 joint locations per human, following the COCO [45] pose format. Additionally, the dataset provides SMPL parameters through the application of the EFT [25] (Exemplar Fine-Tuning) fitting method, further enhancing its utility.

Figure 2.3. An example image of OCHuman dataset.

Other markerless datasets are Mirrored-Human [46], HUMBI [8], THUman 2.0 [10], SSP-3D [47], MultiHuman [11], LSP-e [38], [48], PoseTrack [49], 3DOH50K [5], UP-3D [50], MPII [51], MPI-INF-3DHP [37] (MuCo-3DHP) [16], MannequinChallenge [52], ZJU-MoCap [53], and Mimic The Pose [54]. Table 2.2 shows the list of markerless datasets mentioned.

The Mirrored-Human dataset includes Internet videos of people standing in front of mirrors, providing insight into geometry and depth perception. PoseTrack is a dataset for estimating the 2D pose of multiple individuals. The HUMBI dataset was collected in a laboratory environment with 107 cameras using multi-view reconstruction, while ZJU-MoCap used a similar approach with 21 cameras. THUman 2.0 contains 500 3D scans of humans with SMPL parameters fitted to these scans. MPII is another dataset with 3D annotations. Another markerless dataset, 3DOH50K, was

specifically collected for object occlusion with six different viewpoints. SSP-3D comprises a limited number of in-the-wild images where VIBE is used to initialize SMPL parameters, which are then fine-tuned using its optimization method. MultiHuman also uses a similar approach but employs multi-view footage of people with object occlusions. The UP-3D dataset utilizes a method akin to the unsupervised version of SMPLify to acquire SMPL parameters. LSP and its extended version, LSP-e, contain 2D human poses in natural settings. Mimic The Pose is a human dataset that was specifically collected for self-contacting humans. The MannequinChallenge dataset was collected from the viral internet video trend in 2016. In this trend, people remain motionless while the cameraman roams around the room, providing a stable environment for the scene and objects while the camera is moving, allowing researchers to apply structure-from-motion techniques. MPI-INF-3DHP is a 3D pose dataset collected using multi-view cameras in a green background studio to replace the background with in-the-wild scenes. The authors further improved the dataset in MuCo-3DHP as a data augmentation. 3DOH50K utilizes a method akin to the unsupervised version of SMPLify to acquire SMPL parameters.

In 3D HPS methods, researchers often use datasets that go beyond 3D human models. Researchers commonly incorporate datasets that provide 3D pose annotations to improve training, given the intrinsic connection between 3D HPS and 3D pose estimation. Furthermore, researchers often report their pose prediction performance on dedicated benchmarks, such as CMU Panoptic [1], to demonstrate the versatility and efficacy of their models. Additionally, datasets containing silhouettes are invaluable for the shape estimation component of deep learning architectures. Researchers aim to develop more comprehensive and robust 3D HPS methodologies capable of addressing the multifaceted challenges of this task by encompassing a variety of data sources. Moreover, the integration of multimodal data sources, such as depth maps, RGB images, and even textual descriptions, holds promise for advancing the state-of-the-art in 3D human pose estimation. By continually pushing the boundaries of data utilization and model architecture design, researchers strive to unlock new avenues for improving the accuracy, robustness, and real-world applicability of 3D HPS systems.

Table 2.2. Markerless dataset comparison table.

Name	# Frames	SMPL	In-the-wild
Mirrored-Human	1.8M	Mirrored-Human	Yes
ZJU-MoCap	1500+	EasyMoCap	No
HUMBI	17.3M	HUMBI	No
MPII	25K	EFT	Yes
PoseTrack	66K	EFT	Yes
MannequinChallenge	25K	SMPLy	Yes
3DOH50K	50K	3DOH50K	No
SSP-3D	311	SSP-3D	Yes
OCHuman	4731	EFT	Yes
THUman 2.0	500	THUman2.0	No
UP-3D	8.5K	EFT	Yes
MPI-INF-3DHP	1.3M	SPIN	No
Mimic The Pose	3.7K	Mimic The Pose	Yes
LSP-e	10K	EFT	Yes
MuCo-3DHP	200K	SPIN	No

2.2.3. Simulation Based Datasets

Obtaining 3D surface annotation of people in natural scenes is a challenging task. However, with improving simulations that provide more realistic and visually appealing renders, the gap between in-the-wild and simulation datasets is narrowing each year.

Synthetic datasets, created through computer-generated environments and digital human avatars, offer several critical advantages. Synthetic datasets provide a controlled and reproducible setting, free from real-world limitations such as lighting variations, occlusions, and noisy sensor data. This enables researchers to test and fine-tune algorithms with precision. Moreover, synthetic datasets offer the invaluable benefit of ground-truth annotations, including accurate 3D pose and shape information,

thus serving as a gold standard for model training and evaluation. Additionally, these datasets can be customized to fit specific scenarios and requirements, making it easier to explore a wide range of challenging conditions that may be difficult to replicate in real-world settings.

AGORA (Avatars in Geography Optimized for Regression Analysis) [29] is a dataset collected via simulation that presents highly realistic synthetic humans in six different scenes. The authors commercially purchased 3D scans of people and placed them in scenes with high-resolution textures and different body poses. The authors obtained SMPL/-X parameters of the bodies by fitting them to scans of 4240 individuals, including 257 children. By adding various poses of individuals, AGORA includes 173K distinct person crops, with 14K train and 3K test frames, each containing five to fifteen people. Furthermore, by refining previous techniques with AGORA, they demonstrate its applicability to real-world inputs by achieving higher accuracy on 3DPW. The authors of AGORA recently published a new and expanded version, similar to AGORA, called BEDLAM [32], which includes motions.

The SURREAL [55] dataset is simulation-based and renders bodies created with the SMPL body model on random background images. Similarly, the GTA-Human [35] dataset utilizes the popular video game GTA-V's game engine and fits SMPL annotations to the images.

2.3. Techniques

Significant effort has been dedicated to the field of 3D HPS estimation. Although the number of publications explicitly targeting occlusion is relatively limited, these endeavors have drawn inspiration from the broader body of previous works. Therefore, analyzing the evolution of 3D HPS can offer valuable insights for developing occlusion-robust solutions. Most of the methods are inspired by the previous efforts done in pose estimation algorithms either in 2D or 3D. Also, studies on the image level understanding like silhouette or instance segmentation are also useful in this context.

Figure 2.4. An example image of AGORA dataset.

2.3.1. Human Pose and Shape Estimation

One commonly used taxonomy in this field categorizes approaches as either top-down or bottom-up. Top-down approaches involve human detection within the scene as the initial step and treat the problem as a single-person reconstruction task. In contrast, bottom-up approaches begin by identifying the joints or limbs of all individuals in the image and then merging them to construct complete human body representations. Although the trend is on the bottom-up versions, we investigate both approaches.

A year after introducing the well-known SMPL model, SMPLify [6] became a prominent regression model. Its primary objective is to identify the 2D joints of the human body from a single image, which is often facilitated by an external pose detector. Subsequently, a projection-based approach fits these detected joints to the SMPL model. It should be noted that SMPLify's effectiveness is dependent on the presence of only one individual in the image. Additionally, it requires pose priors derived from the SMPL pose space to handle occluded body parts effectively. The loss functions on shape parameters use synthetic datasets to use shape parameters when necessary.

Another notable work, HMR [4], extended SMPLify and redefined the approach by developing an end-to-end framework for regressing SMPL parameters from a single image. HMR iteratively regresses 3D pose and shape parameters based on reprojection, ensuring a more holistic estimation process. To verify the accuracy of their regressed human body, a discriminator is employed to determine whether the estimate corresponds to an entry in the SMPL database. However, it is important to note that the effectiveness of HMR depends on the person who is centrally positioned in the image and does not address challenges related to occlusion.

In the year of HMR's introduction, Zanfir et al. [56] developed an approach to regress the DMHS [57] model from an image and convert it to the SMPL model. They also introduced an adaptive occupancy avoidance system that considers the distance between two reconstructed models, contributing to a more accurate estimation. The authors then extended their work by providing an integrated bottom-up pipeline for multi-person localization, pose, and shape estimation. The approach involved scoring the potential limbs of individuals in the image and constructing the pose of each person through a binary integer program. Subsequently, the authors regressed the SMPL parameters using an encoder and decoder architecture. It is important to note that the algorithm, while capable of simultaneous multi-person detection, operates relatively slowly due to the complexity of the integer programming component.

Pioneering works such as HMR and SMPLify excel at recovering images featuring solitary individuals. However, Kolotouros et al. [18] extended this approach to accommodate multi-person scenarios by introducing their architecture, SPIN. SPIN combines SMPLify with a person localization module, enabling the simultaneous estimation of multiple individuals within an image. The SPIN method first localizes each person using an object detector and then iteratively regresses the SMPL parameters in a loop. It is important to note that this innovative framework may face challenges when dealing with occlusions, as it was primarily designed to handle unoccluded scenarios. Note that this approach was quite popular until the introduction of occlusion-robust estimators and is still in use for trivial samples.

The GraphCMR [58] approach extends the effectiveness of Graph Convolutional Neural Networks (GCN) [59] for object mesh reconstruction to the domain of human recovery. This method encodes image features and utilizes a GCN to reconstruct the human mesh in a coarse-to-fine fashion, demonstrating an ability to handle object and self occlusions effectively. However, it is important to note that GraphCMR is primarily focused on single-person reconstruction and does not offer capabilities for multi-person reconstruction or handling scenarios involving person-to-person occlusion.

VIBE [7] is another popular method for synthesizing the motion of people in 3D from videos. The architecture involves extracting features from consecutive frames using a CNN backbone independently. The extracted features are then associated using GRU [60]. Finally, the motion is generated from a Generator while a discriminator trained on AMASS [44] database decides whether the produced sequence is realistic or not, similar to HMR.

GLAMR [12] aims to recover human pose and shape in dynamic environments from videos. They use a generator to fill in missing poses due to occlusion or truncation and adopt a trajectory estimator to track the position of people in a common coordinate frame. Note that their approach includes an occlusion-aware mechanism where moving dynamic cameras are supported, too.

2.3.2. Occlusion Robust Studies

Strategies for addressing human occlusion exhibit notable diversity, reflecting the inherent complexity of this problem. However, recent research has established a set of standardized practices. Contemporary studies either explicitly address occlusion or evaluate their methods on occlusion-centric datasets. However, existing surveys on human pose estimation and mesh reconstruction have not comprehensively addressed the crucial aspect of occlusion benchmarking. This work aims to fill this gap by categorizing, summarizing, and critically evaluating diverse occlusion-handling methods from various perspectives.

To enhance robustness against occlusions, a simple yet effective approach is to use data augmentation in the training dataset. Huang et al. [61] proposed a data augmentation method in one of the earliest studies focusing on human body reconstruction under occlusion. They treated each occluded test sample as a sparse linear combination of unoccluded training samples. However, it is important to note that the study primarily dealt with silhouettes. Sarandi et al. [62] introduced another augmentation strategy by incorporating object segments from the PASCAL VOC [63] dataset and geometric shapes such as circles and rectangles to serve as synthetic occlusions in the Human3.6M [33] dataset.

Rafi et al. [27] conducted a pioneering effort in addressing the reconstruction of occluded body parts. Their approach leveraged joint detection facilitated by the Kinect 2 SDK [64]. To address the limited performance of the Kinect 2 SDK (Software Development Kit) in handling occluded joints, they devised a regression forest to estimate the probability of pixel regions containing a joint. Additionally, their method incorporated semantic reasoning concerning the occluding objects to enhance estimation accuracy. It is important to note that their algorithm only applies to depth images and is limited to scenarios involving a single individual.

Another technique that is commonly used involves guiding the feature extraction process with additional input channels. For example, Zhang et al. [5] utilized information from a coarse saliency map and transformed the problem into a UV-map inpainting challenge. Their approach includes a sub-network that generates a partial UV map representing the visible parts of the person. The image features and the saliency map are jointly used to complete the missing regions. This method is effective in recovering a single person who is occluded by objects, but it may not be able to handle scenarios where occlusion is caused by other individuals.

In addition to providing guidance at the image level, the design of the loss function is crucial in the 3D HPS estimation process. Jiang et al. [65] introduced an innovative approach by directly regressing SMPL models from images using a modified RCNN [66]

model, which they named SMPL RCNN. They also incorporated an interpenetration loss, which represents the human surface as an inverse signed distance function. This enabled the detection of vertices from one person penetrating the surface of another. Furthermore, a depth order loss was utilized to ensure alignment between the projected meshes and the instance segmentation masks of the individuals, thereby improving the accuracy of their 3D pose and shape estimation.

Zhou et al. [24] used saliency maps to address occlusions and represented the human subject within an amodal mask, even when partially occluded. They used two concatenated hourglass modules to complete the invisible parts of the individual effectively. Finally, they introduced a discriminator to assess the reasonableness of the generated mask. Additionally, they included an attention network to retrieve the content of the obstructed person, enabling precise estimation of 3D pose and shape by fitting the SMPL model.

In addition to conventional methods such as bounding boxes and cropping, alternative localization techniques have been explored. One such technique is using a grid structure. BMP [23] adopts a grid-based approach, where the image is divided into a grid, and the probability of a person's presence is calculated for each cell. It is worth noting that their grid structure is 3D, which allows it to capture the depth order of multiple individuals within the scene. Additionally, the authors expanded their dataset by superimposing random objects onto the pose keypoints of the individuals, thereby increasing the dataset's diversity and the robustness of their model.

ROMP [17] is a bottom-up approach for addressing occlusion. It localizes each person within a body center heatmap and employs another parameter map to regress SMPL and camera parameters across a grid structure. The parameter map is obtained by sampling it from the body center heatmap by interpreting it as a 2D probability density function. To address occlusion challenges caused by overlapping body centers, ROMP introduces a repulsion function that pushes the centers apart within the heatmap. This ensures the centers do not overlap and maintains the data integrity.

PARE [28] utilizes an attention mechanism [67] to associate body joints that may be separated by occlusion. The authors conduct an occlusion sensitivity analysis by predicting body-part guided attention masks. Their architecture distinguishes the body segments in 2D and extracts 3D features concurrently. Then, a Part-Attention mechanism fuses the extracted features to predict SMPL parameters.

OCHMR [22] expands on the concept of heatmap localization by utilizing feature extraction through a unique approach inspired by Batch Normalization [68]. After localizing the human within the heatmap, OCHMR introduces context normalization layers (CoNorm) to regress shift and scale parameters for the extracted features. The process of feature extraction proceeds through the CNN layers after the shift and scale adjustments. OCHMR, like ROMP, introduces random noise to the heatmap centers, which allows it to tolerate body-centered occlusions and reduces noise from heatmap generators. However, it is important to note that OCHMR necessitates a separate run for each person in the image, distinguishing it from other methods that handle multiple individuals in a single pass.

Recent studies have focused on improving robustness to person-to-person occlusion in crowded scenes, such as family or souvenir photos. One notable example is 3DCrowdNet [3], which introduces a bottom-up crowd reconstruction framework for direct SMPL parameter regression for individuals within the image. In this approach, after localizing a person in the crowd, an off-the-shelf pose detector is utilized to generate a heatmap of joints. By combining the heatmap with image features, the authors use a Graph Convolutional Neural Network (GCN) to regress the SMPL parameters. They also evaluate their method on subsets of standard datasets chosen by CrowdIndex [69] that are representative of crowded scenes. However, like other approaches, 3DCrowdNet also requires a separate run for each person in the image.

The authors of ROMP expanded their approach to address crowded scenes, including modeling infants and babies using the Bird-Eye-View (BEV) [15] method. BEV is a framework that recovers individuals from a top-down perspective, aiming to miti-

gate depth ambiguity in crowded scenes. To reconstruct infants, the authors leverage the relationship between a person's height and age in the image space. Depending on the observed height, individuals are represented as a convex combination of SMPL and SMIL [70], which is a variation of SMPL specifically designed for infants and babies. This approach enables a more precise and age-appropriate reconstruction in crowded settings.

ImpHMR [71] applied neural view synthesis techniques to address the issue of human occlusion. Their objective is to reconstruct a person from different viewing angles, including self-occluded body parts.

3. RELATED WORK

This section examines in detail the experiments conducted using state-of-the-art approaches in both deep learning methods and datasets. Additionally, the method used to quantify occlusion in datasets is criticized, and its drawbacks are revealed with examples.

3.1. ROMP: Monocular, One-stage, Regression of Multiple 3D People

A common approach to estimating multi-person 3D human pose is to localize each person by detecting their bounding boxes and running a single-person reconstructor consecutively on each cropped image. However, this approach is not effective when dealing with occlusion, as occluded and occluding people share similar regions, even if the localization step is accurate (which is often not the case). Therefore, methods that do not differentiate between humans and other objects gained popularity due to their ability to recover humans who are spatially close to each other. However, these methods use a single-stage localization approach.

ROMP, on the other hand, employs a bounding-box-free localization and SMPL parameter estimation framework that works simultaneously. The feature extractor backbone produces three maps on the image: the first one locates people as a body center heatmap, the second one regresses the SMPL parameters at each cell on a grid, and the third one estimates the camera parameters for each cell. The second and third maps are concatenated, and ROMP samples from the cells of those grids based on the probability of a person being in that cell, interpreted from the body center heatmap as a 2D probability distribution. Prior to sampling, the local maxima of each body center are repelled from one another to highlight the differentiation between two agents in a two-dimensional space. This Collision-aware representation prevents the multiple reconstructions of the same person. This approach is also utilized in the successor methods in various versions for handling severe occlusions.

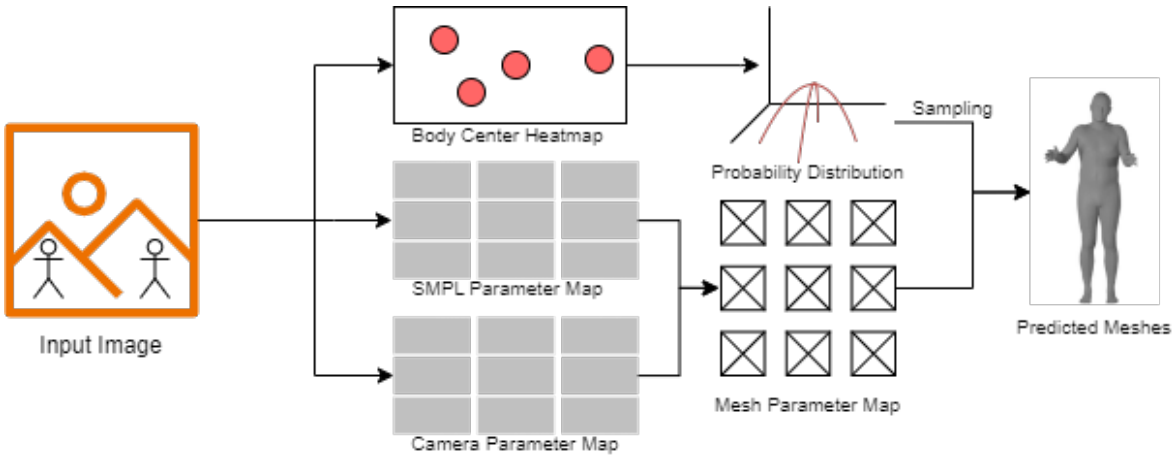

Figure 3.1. ROMP Architecture.

3.2. BEV: Monocular Regression of 3D People in Depth

Crowd images are a type of multi-person footage that presents unique challenges. Methods designed to address these challenges can be applied to other multi-person tasks, including 3D HPS. Souvenir and family photos are useful for testing occlusion robustness due to heavy occlusions.

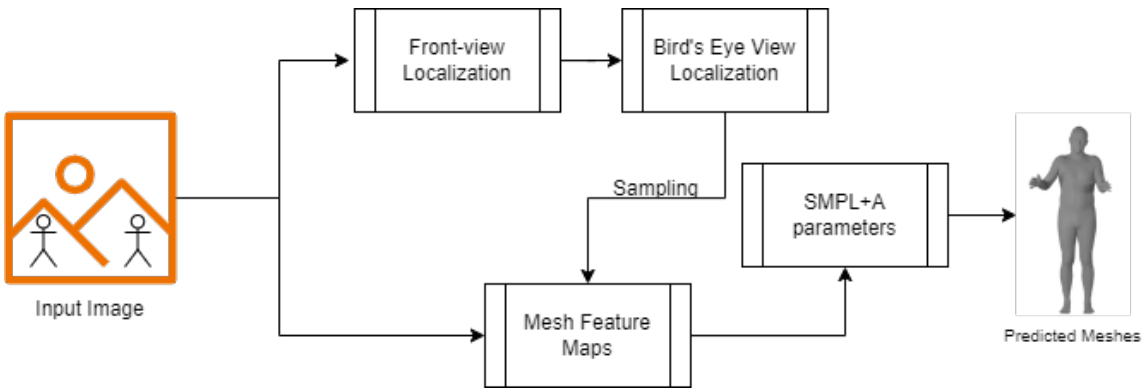

Figure 3.2. BEV architecture.

The BEV (Bird's-Eye-View) algorithm can be used to regress a human's 3D surface and pose in crowd scenes. The authors aim to accurately recover a bird's-eye-view of the scene to explicitly handle the problem of depth ambiguity. Additionally, they estimate the age of each person in the image beforehand to determine the relation between depth and height. They use an appropriate model for the predicted age period,

representing each age as a convex combination of SMPL [40], an adult model, and SMIL [70], a child model. Therefore, each individual is constructed with an appropriate model to prevent issues such as mistaking children for distant adults.

3.3. CrowdIndex

CrowdIndex is a metric used to quantify crowding in an image. It was introduced with the CrowdPose [69] dataset. The main motivation behind the CrowdIndex is to highlight that the primary challenge in human-related vision tasks is not the number of people in the scene, but rather the occlusion caused by the crowd. This metric requires the bounding boxes and 2D keypoints of the joints for each person. The number of joints belonging to the ith person, denoted as N_i^a, and the number of joints not belonging to them, denoted as N_i^b, are counted within their bounding box. The *Crowd Ratio* for each person is calculated as N_i^b/N_i^a, and the *Crowd Index* for a frame is the average of the crowd ratios of all people in the image. The formula is

$$\text{Crowd Index} = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i^b}{N_i^a}, \tag{3.1}$$

where n is the number of people in the frame.

Figure 3.3 illustrates the formulation using an example. In Figure 3.3a, the keypoints of the person on the left are marked in red, and the joints belonging to the other person but falling inside the bounding box of the person on the left are marked in blue. Therefore, by counting the number of blue and red points, which are 5 and 24, respectively, we can conclude that the Crowd Ratio for the person on the left is $\frac{5}{24}$. Figure 3.3b shows the person on the right with a Crowd Ratio of $\frac{7}{24}$.

The authors of Crowd Index demonstrated that their dataset is more uniformly distributed in terms of occlusion. Some studies [3] aimed to demonstrate their robustness to occlusion by typically testing their methods on subsets formed by the Crowd Index. However, the metric has several drawbacks that make it not suitable for occlusion quantification in human datasets.

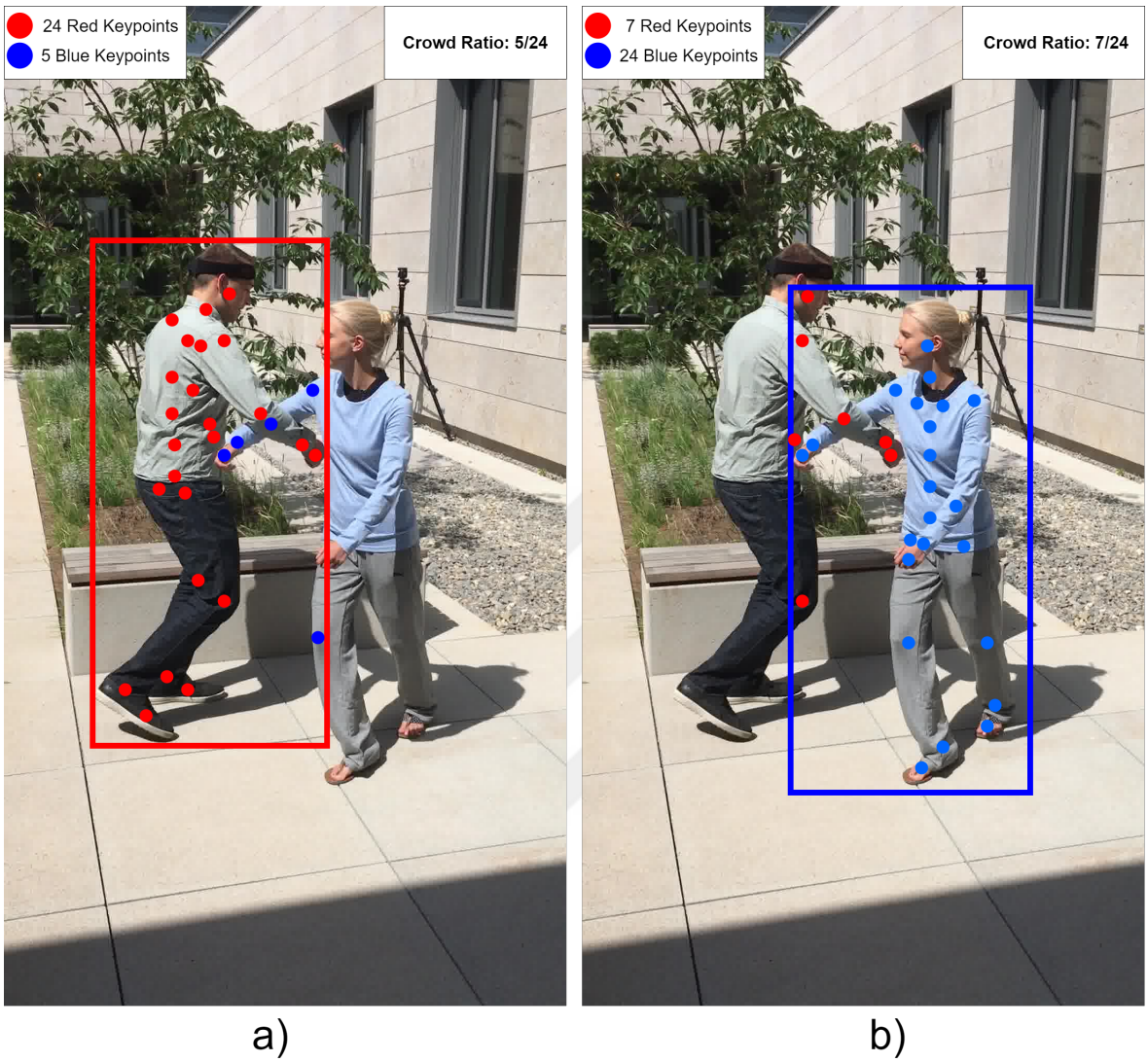

Figure 3.3. Crowd Index for the person on the left(a) and right(b).

The Crowd Index can not differentiate between the person being occluded and the person occluding them. This is because the bounding boxes do not indicate a depth order between the humans. Thus, the metric applies the same approach to both parties, which may lead to a higher occlusion report for occluding individuals compared to occluded ones. Figure 3.4 shows a person at the front who is fully visible and another person at the back who is partially obscured. The Crowd Index, however, calculates higher values for the person in the front with a high margin due to not distinguishing between the occluding and occluded people. It should have returned a near-zero value for the person in the front and a relatively high value for the person in the back due to partial occlusion.

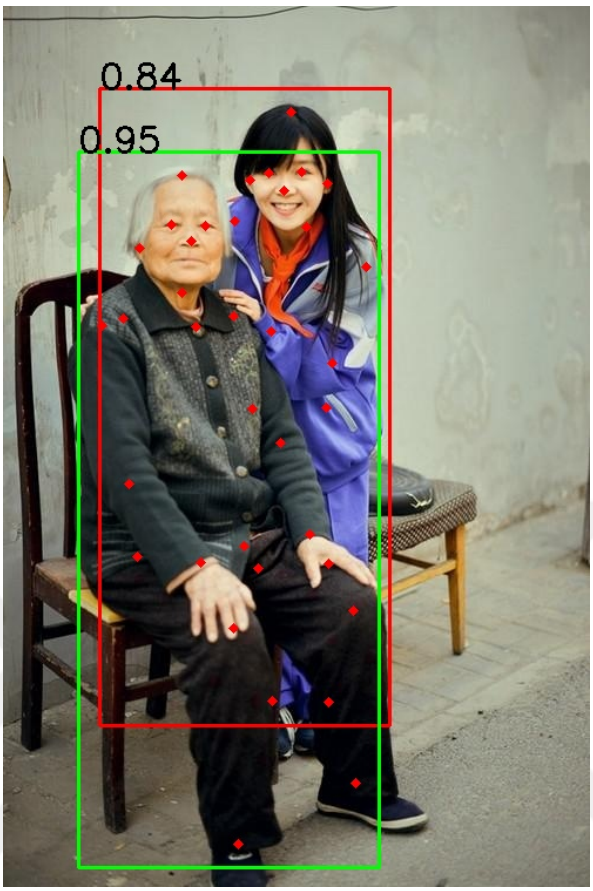

Figure 3.4. Failure of Crowd Index for occluding person.

Secondly, the Crowd Index accumulates based on the number of people occluding a person, regardless of the visibility of the occluded person. Consider a scenario in which multiple individuals obstruct a person's right arm. It will not be visible no matter how many people are blocking it. However, the Crowd Index still increases the occlusion value for the occluded person as the number of people occluding them increases. An example failure of Crowd Index due to counting occlusion for each person occluding is shown in Figure 3.5. The right shoulder of the person at the back is occluded by both people at the front. However, although the visibility of that joint remains the same, the index of the person at the back increases as if the visibility of the person changes. Furthermore, the same joint contributes to occluding two people. This means that the Crowd Index has no upper limit, and as the number of occluding individuals increases, the occlusion value approaches infinity. However, the occlusion of a person in 2D can be expressed as a percentage.

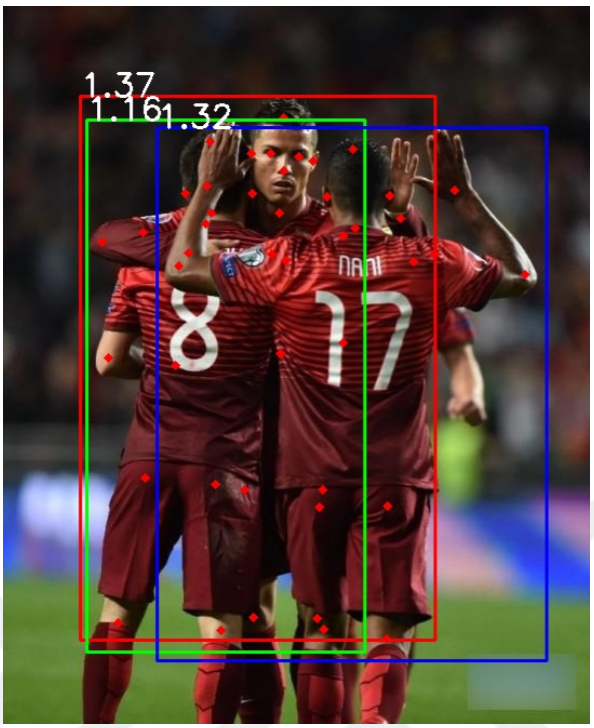

Figure 3.5. Failure of Crowd Index for repetitive occlusion.

One issue with the current approach is the imprecise bounding boxes. The bounding box for a person is drawn as the smallest rectangle that contains all of their keypoints. As a result, a person who is spreading their arms and legs will have a large bounding box, even though they only occupy a small number of pixels. However, the Crowd Index does not distinguish between a person and their bounding box. Therefore, if another person's joint projection falls within the bounding box, but not on the person themselves and is not occluded, the Crowd Index counts it as occluded, even if it is visible. An example failure of the Crowd Index due to relying on bounding boxes is demonstrated in Figure 3.6. Although the person on the right suffers from minimal occlusion, the one on the left is completely visible. However, the Crowd Index does not differentiate between a person and its bounding box, resulting in occlusion being assumed for both individuals. This includes the person on the left, who has zero occlusion but is still given a higher occlusion value.

In summary, the Crowd Index has been found to produce false positives, have low precision, and report occlusion levels higher than they actually are. This drawback

leads to inaccurate filtering of samples in the dataset, and the subsets formed by it do not provide a challenging occlusion benchmark. Previous studies that reported their occlusion robustness performance on the subsets created using the Crowd Index may not have achieved the level of robustness they assumed.

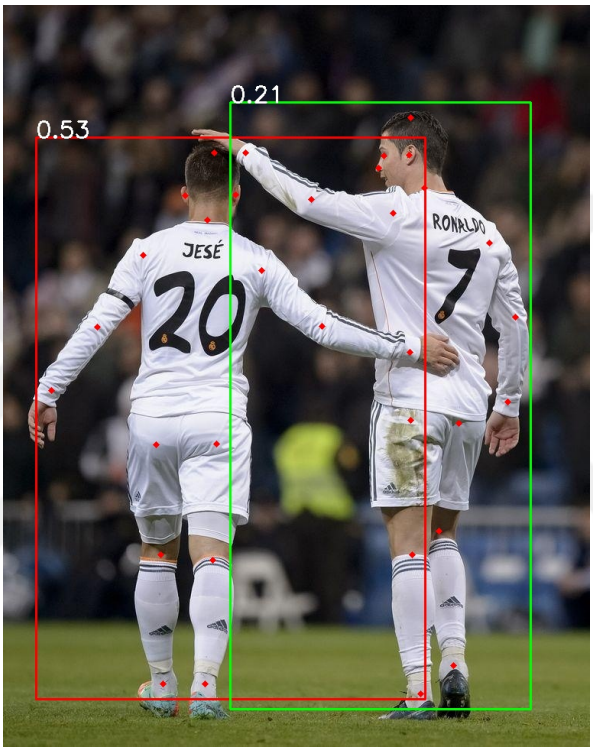

Figure 3.6. Failure of Crowd Index due to bounding boxes.

4. METHODOLOGY

The purpose of this study is to quantify occlusion for each person in a given image and differentiate and rank images based on their degree of occlusion. The research is carried out in several stages. First, a segmentation-based index that is an alternative to the Crowd Index is proposed by addressing the limitations of the Crowd Metric. In the second stage, we extend our work to a fine-grained version of it that depends on the number of occluded pixels by eliminating the need for keypoints and producing our own segmentation mask by rendering the 3D model surface to the image plane. Then we introduced additional properties to our index that allow us to categorize occlusion based on the occluded body segment. Following that, we designed a framework that allows us to weigh each body segment's occlusion and form the most challenging subset by learning the weights from methods. We also investigated the effect of body orientation and validated our occlusion quantification approach in the self-occlusion cases. In addition to measuring occlusion, we revised the commonly used evaluation metric to place more emphasis on reconstructing visible joints rather than evaluating the entire body, which is possible thanks to our novel index, and suggested a new metric for occlusion robust methodologies.

4.1. A Novel Occlusion Index

Our proposed index aims to measure occlusion in each frame and generate a subset that specifically focuses on occlusion for datasets. The initial stage addresses the primary constraints of the Crowd Index. The subsequent stages improve the proposed index with various components and allow for a more in-depth analysis of the innate features of occlusion. The improved and fine-grained version of our proposition is introduced in Section 4.1.2 using only the minimum labels and proposing a pipeline. In Section 4.1.3, we labeled each body segment and investigated their occlusion effect separately and the body's root orientation. Lastly, in Section 4.1.4, we proposed a framework that weights the importance of each body segment in terms of visibility.

4.1.1. Occlusion Index

Although the Crowd Index is able to detect occlusion for trivial cases, as demonstrated in Section 3.3, there are some obvious drawbacks that can lead to inconsistent occlusion scores and an increased number of false positives. Before delving into the fundamental problems that cause this inadequacy, let us describe our proposed *Occlusion Index* (OI) [39].

Our proposed index requires individual segmentation masks and poses of people in an image. For each pose of a person in the image, the proportion of its joints falling on another person's mask is calculated. The set of keypoints that do not fall onto the associated bounding box is

$$k_i = \{k_i^j | k_i^j \notin M_i \wedge k_i^j \in K_i\},\ j = 1, ..., m, \tag{4.1}$$

where m is the total number of joints, K_i is the set of all joints of ith person (meaning $|K_i| = m$), k_i^j is the jth joint, and k_i is the subset of joints that do *not* fall on the ith person's mask M_i. The Occlusion Index for the ith person is

$$o_i = \frac{|k_i|}{|K_i|}, \tag{4.2}$$

where $|k_i|$ is the cardinality of the set of occluded joints and $|K_i|$ is the cardinality of the set of all joints belonging to person i. The occlusion score for a frame is the arithmetic mean of the occlusion scores for all people in that image

$$o_{frame} = \frac{1}{n} \sum_{i=1}^{n} \frac{|k_i|}{|K_i|}, \tag{4.3}$$

where n is the number of people in the frame.

Figure 4.1 shows an example calculation of the Occlusion Index. In Figure 4.1a the person on the left, denoted by a red mask, is completely visible, and the corresponding joints are highlighted as white keypoints. On the other hand, 4.1b visualizes the Occlusion Index for the person on the right. The person on the left occludes the right elbow of the person on the right, marked with a red keypoint. Therefore, the Occlusion Index for the person on the left is 0, and for the person on the right, it is $\frac{1}{24}$.

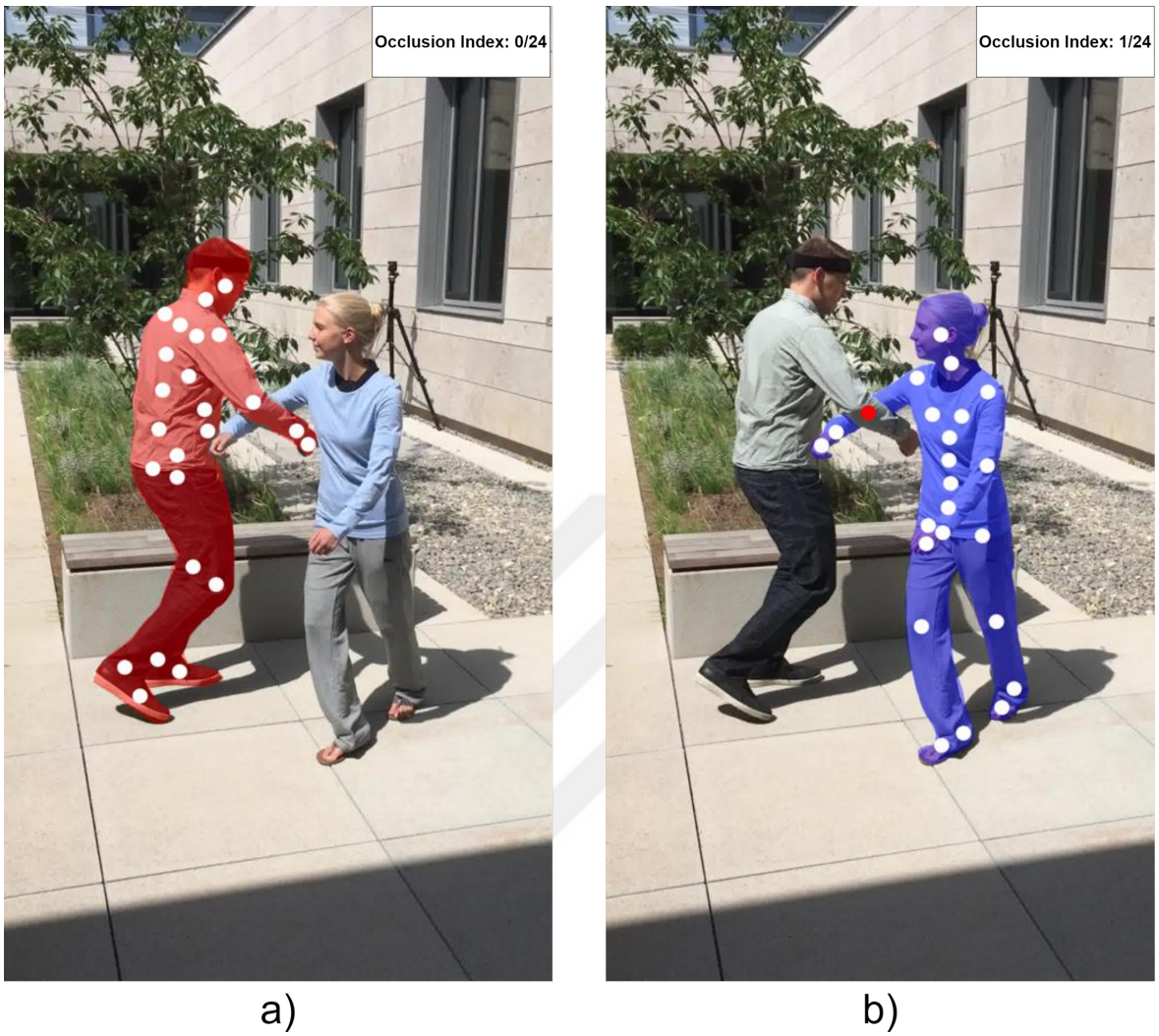

Figure 4.1. Occlusion Index for the person on the left(a) and right(b).

The Occlusion Index uses a pose-based calculation to evaluate the joints in each image against the mask to which it corresponds. A joint is considered visible if it falls within the boundaries of the same mask it represents. Conversely, a joint is considered occluded if it does not fall within the corresponding mask. In that case, it either falls into another person's mask or on an object without a mask. It is important to note that the Occlusion Index is calculated per joint, not per mask.

Alternatively, an area-based calculation, similar to the Crowd Index, could be considered. Such an approach would require examining all the joints falling into that segmentation mask and classifying whether each keypoint belongs to the same person. It is important to note that the Crowd Index is computed per bounding box or mask

if instance-segmentation masks would have been used. Also, given the fixed number of keypoints for each individual, the Crowd Index can be simplified by examining the number of joints that do not belong to that individual but still fall within the individual's bounding box. However, in this case, one would have needed the masks of the occluded regions as well.

The masks for the occluded regions are not needed for the pose-based index, such as the Occlusion Index. The index is computed by determining whether each joint falls on its own mask or not. This makes it compatible with commonly used human instance segmentation models, as they label only visible pixels. Area-based indices, similar to the Crowd Index, require masks for the occluded regions of the individual. This is because they rely on binary classification of all joints that fall on that region of pixels. This limitation makes them incompatible with commonly used segmentation models.

Due to the aforementioned properties, the Occlusion Index does not have the Crowd Index's limitations described in Section 3.3. Let's review each example and demonstrate the superiority of the Occlusion Index.

4.1.1.1. Pseudo Depth Order. The ultimate superiority of the Occlusion Index over the Crowd Index is its ability to distinguish occluders from occluded persons. In Figure 4.2a, the Crowd Index produces a higher value for the person in the front, even though this person is not occluded by anyone else. The bounding boxes cause this confusion. Since no depth order can be defined between bounding boxes, the Crowd Index only considers the 2D spatial overlap between a bounding box and counts the keypoints in the box, regardless of whether those are occluded. The Occlusion Index, on the other hand, uses instance segmentation masks that highlight only the visible regions of a person. Therefore, the regions that are not highlighted by a person's mask but contain joints of that person can be assumed at the back. In corollary, it can be determined whether a person's keypoints are in the very front (visible), which is our Occlusion Index's main driving factor. This feature allows us to distinguish

keypoints of a person based on whether they fall within the mask of that person or not, corresponding to visible or occluded, respectively. Therefore, in Figure 4.2b, the Occlusion Index produced a zero occlusion score for the person in the front and a relatively high value for the person partially occluded. The reason is that all the person's keypoints denoted by the green instance mask fall into the green regions. However, some of the keypoint of the person at the back who is partially occluded fall onto the green regions, which emits an occluded signal for that keypoints. Although this is an obvious improvement of the previous method, this approach also has some drawbacks that need to be addressed. Determining a region's occlusion using a single keypoint may be misleading and an under-statement of the problem. However, this study investigates such deficiencies in the following sections.

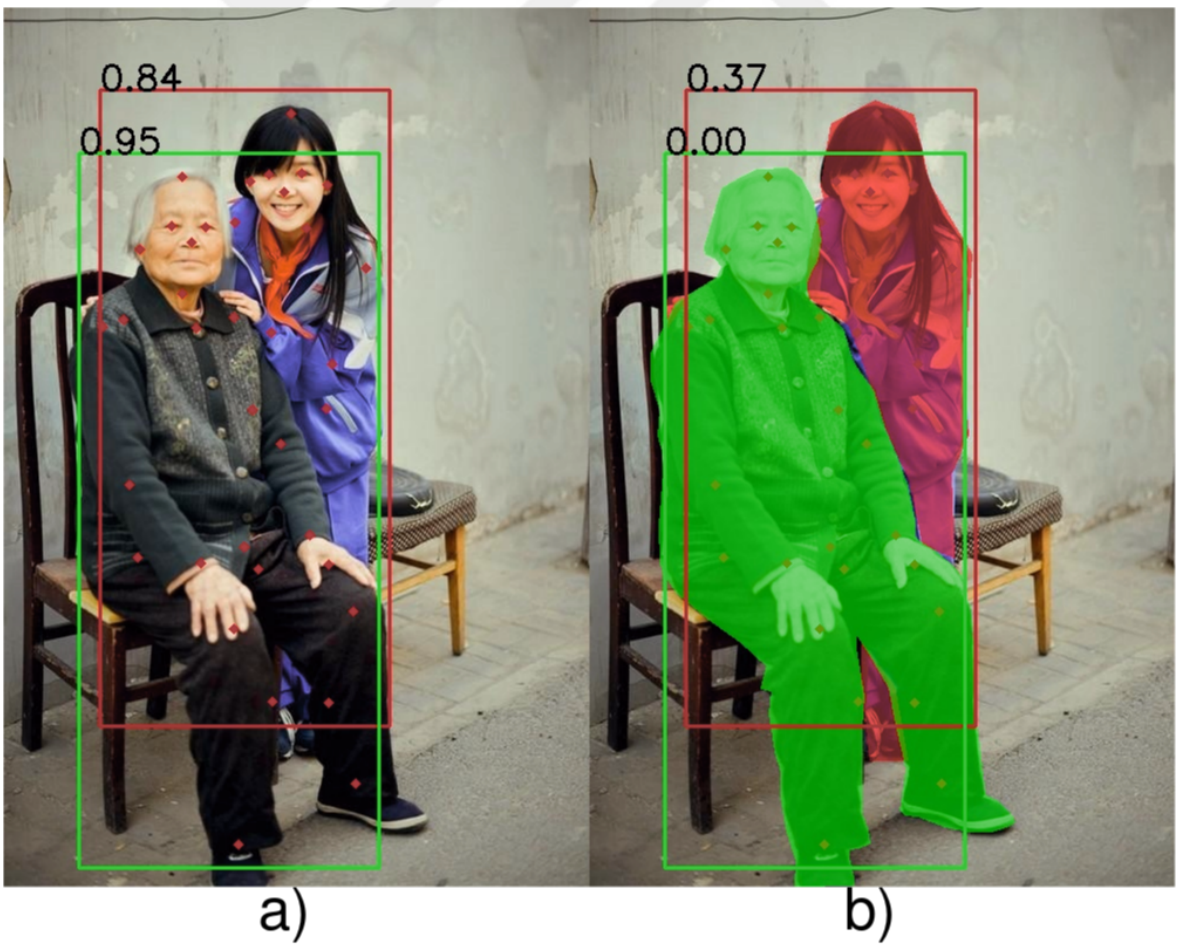

Figure 4.2. Crowd Index (a) does not distinguish occluded and occluding people but Occlusion Index (b) has pseudo depth order.

4.1.1.2. Pose Invariant Precision. Another advantage of the Occlusion Index is the score's precision does not change due to spanned poses. The Crowd Index uses bounding boxes; each bounding box is spanned by the most distant keypoints of the occluded person. Thus, the size of the bounding box depends strictly on the pose of the person, although the number of pixels occupied by the person remains the same. The Crowd Index simplifies people to bounding boxes and assumes that the entire rectangle represents the human body in 2D. This property increases the score if there is another person nearby that has some keypoints in the person's bounding box but is neither occluded nor occluding. Consider Figure 4.3a, where the person on the left is fully visible, and the person on the right is minimally occluded. (Actually, he is not occluded by keypoints, but by pixels. This is one of the drawbacks of keypoint-based indices. This study deals with this in the following section.) However, due to the pose of their outstretched arms, they occupy a larger area than they should. This tilts a large area of the intersection of these bounding boxes, and there are keypoints from both parties that fall into the other person's bounding box. Worst of all, the Crowd Index produces a higher value for the person who is fully visible and a lower value for the person who is minimally occluded (or not occluded from the perspective of the keypoints). The Crowd Index depends on the bounding box, and the bounding box depends on the pose. As the difference between the number of pixels occupied by the bounding box and the person increases, the accuracy of the Crowd Index tends to decrease. Therefore, it can be said that the accuracy of the Crowd Index is highly dependent on the pose of the person. In contrast, the Occlusion Index is not affected by pose. Instance segmentation masks only occupy the person's pixels, so they do not degrade accuracy. This claim can be verified in Figure 4.3b. The occlusion score of the person on the left is not affected by the person on the right, as it should be.

4.1.1.3. Invariance to the Number of Occluders. Occlusion is closely related to visibility. If a person's body segment is not visible (assuming it is not truncated and in the frame), it can be counted as occluded. It does not matter how many people occlude that region. However, the Crowd Index does not care about visibility. If a person's right arm is occluded by several people, even if that person's visibility remains the

same, the score produced by the Crowd Index for that occluded person will increase because their keypoints will also fall on the bounding box and increase the total number of keypoints within that bounding box. On the other hand, the instance segmentation masks used by the Occlusion Index produce a binary signal for each pixel, indicating visibility. Therefore, the score produced by the Occlusion Index does not depend on the person or the number of people that make a keypoint invisible. In Figure 4.4, the joints denoted with white points are occluded by both people represented with blue and green segmentation masks. Even if the person in front (with green mask) were not there, the occlusion score for the person at the back (with red mask) would not change.

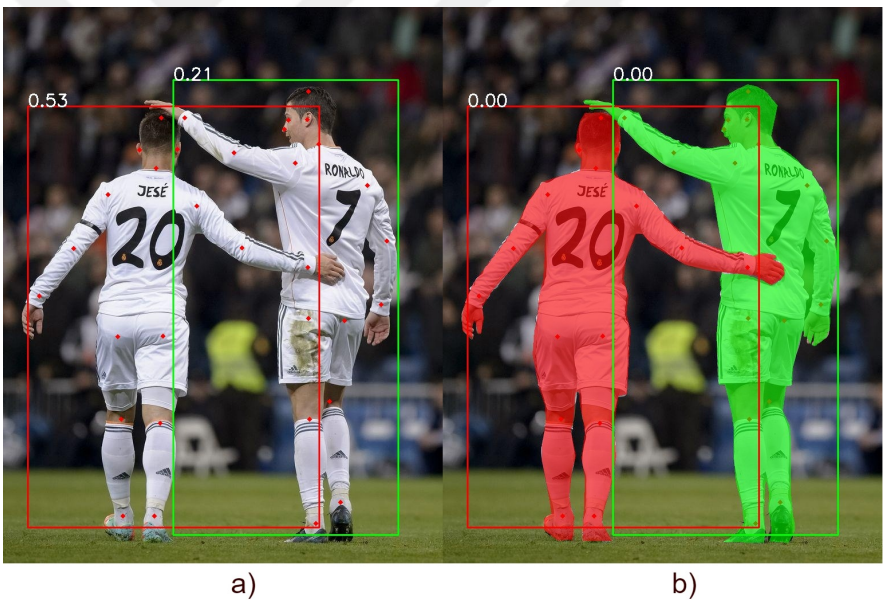

Figure 4.3. The precision of Crowd Index(a) varies depending on the pose, but Occlusion Index(b) is pose invariant. The precision of the Crowd Index tends to decrease, especially for spanned poses.

4.1.1.4. Normalized Score. When comparing two scores, no matter what the context is, having an upper and lower bound for that score is a very useful property. For the Crowd Index, however, there could be an infinite number of people in a person's bounding box. Since the Crowd Index sums the number of all keypoints in that box, this number can increase without an upper limit. (Remember the Equation 3.1). However,

when calculating an occlusion score for a person, the Occlusion Index only considers the visibility of that person's joints, which is a fixed integer. The Occlusion Index creates two subsets for each person's joints; one containing the visible keypoints and the other containing the invisible/occluded keypoints. These subsets are the disjoint partition of the keypoint set, where the union of these subsets is equal to the main set, but the intersection of the subsets is an empty set. Therefore, the Occlusion Index is limited to 0 to 1 because it is defined by the fraction of the number of occluded keypoints, which is less than or equal to the total number of keypoints, and the total number of keypoints, which is a fixed integer no matter what pose format is used. Figure 4.5 demonstrates several examples with different occlusion levels. Remember that the occlusion scores attached to the images in Figure 4.5 are representative. However, the real values are quite close and the order is precisely identical.

Figure 4.4. Occlusion Index is invariant to the number of occluders. If the person with blue color is removed, the occlusion of the white keypoints stays the same.

4.1.1.5. Detection of Object Occlusions. While person-to-person occlusion is a challenging problem, a person can be partially occluded by an object. However, since the Crowd Index only considers the bounding boxes of people and their interaction via keypoints, it cannot detect if an object is occluding a person. The Occlusion Index, on the other hand, includes all cases where a keypoint does not fall on the corresponding instance mask, regardless of whether the instance mask has lost its integrity due to a person or an object. Figure 4.6 is an example of an Occlusion Index detecting object occlusions.

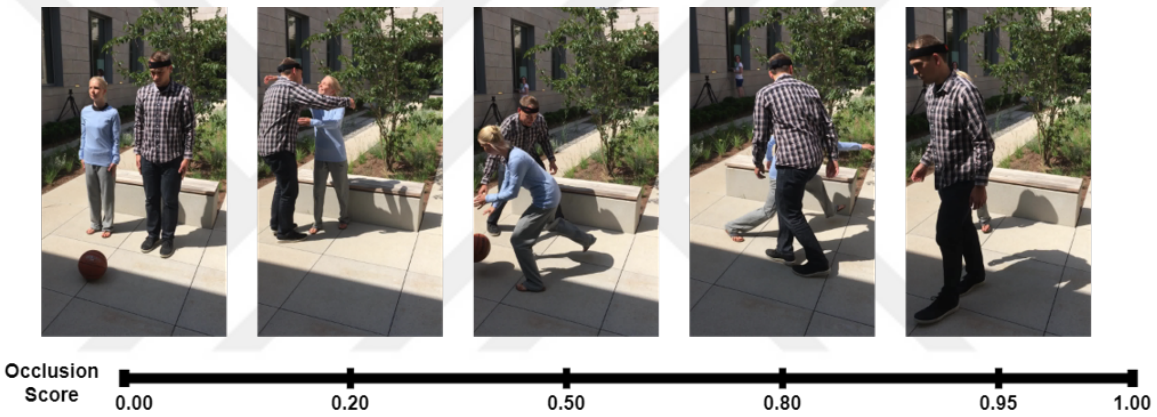

Figure 4.5. Occlusion Index is bounded between zero and one.

For all of the reasons listed in Section 3.3, the Crowd Index tends to report higher occlusion values even for non-occluded people and to produce false positives. However, our Occlusion Index distinguishes between occluding and occluding people and reports a normalized value that is invariant to the person's pose and the number of people occluding. Note also that the Occlusion Index quantifies occlusion not only for 3D HPS problems but also for any task whose solution models may be affected by occlusion, such as human pose estimation and crowd counting.

4.1.2. Regional Occlusion Index

The Occlusion Index allows us to test the performance of the 3D HPS estimators' occlusion robustness to different levels of occlusion. Although it is a great tool to approximate a person's occlusion, it also has drawbacks.

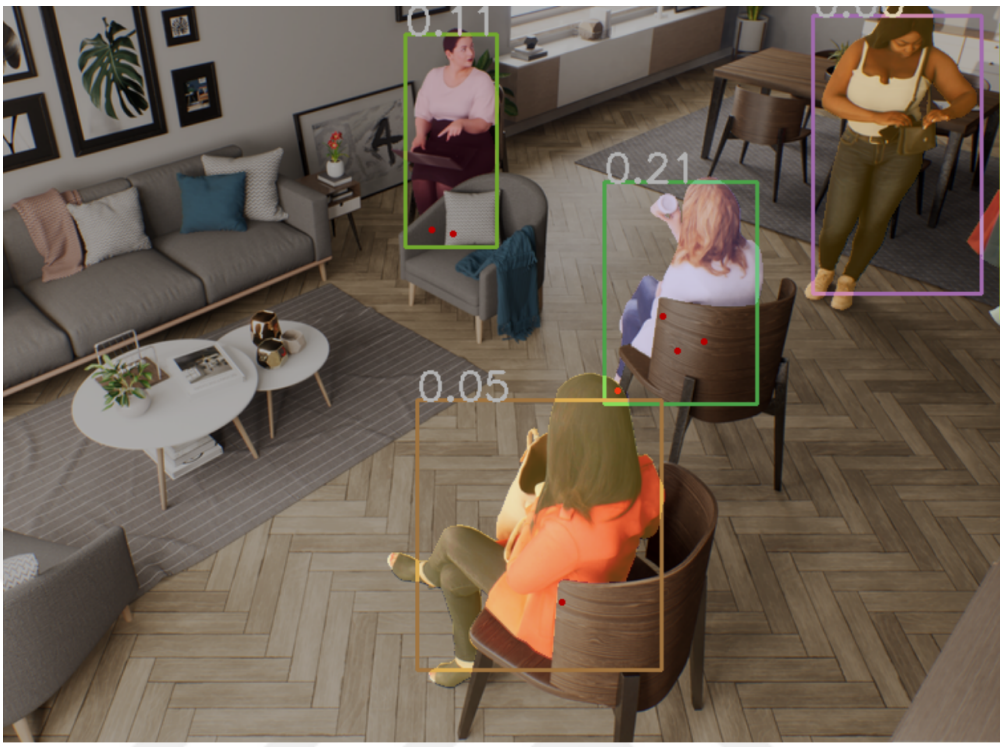

Figure 4.6. Occlusion Index can detect object occlusions.

The Occlusion Index represents each joint by a 2D point. We obtain such points by projecting 3D joint points onto the image plane using a perspective camera model in the context of 3D HPS. However, since each point is represented by only one pixel in the image coordinate frame, each body segment represented by that point is reduced to a single pixel, which oversimplifies the problem. Instead, body segments occupy many pixels, whereas a one-pixel representation can be misleading. Moreover, this simplification leads to binary filtering of each body segment, which causes much information loss.

To overcome this limitation, we decided to look at the ratio of the number of pixels of the occluded regions to the total number of pixels for that person. This makes it possible to quantify occlusion continuously in a more fine-grained manner.

On the other hand, most datasets do not provide the annotation of the instance segmentation masks of occluded regions, which is needed to calculate the total number of pixels where the person resides.

At this point, we prefer to create our own instance masks. Since we are focusing on the 3D HPS problem, all the datasets we work on must provide a 3D human surface model. We projected these 3D models onto the image plane and used these projections as *pseudo-instance masks*.

In Section 4.1.1, compatibility with standard instance segmentation masks is considered an advantage. However, even in this case, there is still a need for occlusion robust instance segmentation models, which also remains an unsolved problem. Rendering 3D human models also eliminates the need for such robust models.

We used the Z-buffer [72] algorithm to create pseudo-instance segmentation masks by rendering 3D human models for datasets without instance segmentation masks. Essentially, we generated the 3D mesh model for each individual using the parametric SMPL model and then projected each mesh triangle onto the image plane. We then determined the depth for each pixel using a triangle intersection test. In cases where a pixel fell on more than one triangle, we assigned its color based on the triangle with the closest depth to the camera, determined by comparing the depths of the corresponding 3D triangles in barycentric coordinates. Figure 4.7 shows the image (a) and the rendered pseudo-instance masks (b).

To derive 2D keypoints, whether visible or not, we projected 3D points provided by datasets corresponding to the SMPL pose joints onto the image plane to estimate the joint positions on the pseudo-instance masks. Maintaining the correspondence between skeletons and mask colors is critical. We also need the 3D keypoint locations of the joint driving SMPL pose. Since we keep track of the occlusion of SMPL surface, the connection between the 3D surface model and the 3D pose format is also important.

The pseudo-masks of the people in the image are obtained as silhouettes regardless of occlusion, as well as visible-only regions. Figure 4.8 shows the input image (a), the visible-only regions (b), and the full rendering of the person at the back as if he were visible (c), and the occluded region of the person at the back (d), separately. The ratio

of the number of colored pixels in Figure 4.8d to the number of colored pixels in 4.8c gives the occlusion in a more fine-grained way, which we call *Regional Occlusion Index*.

Figure 4.7. Input image(a) and rendered pseudo instance masks(b).

The set of silhouette pixels of ith person can be expressed as

$$S_i = \{(x, y) | I_i^S(x, y) > 0\}, \tag{4.4}$$

where I_i^s is the image of silhouette of the ith person (Figure 4.8c) and the set of occluded pixels of ith person is

$$O_i = \{(x, y) | I_i^O(x, y) > 0\}, \tag{4.5}$$

where I_i^o is the image of the occluded region of ith person (d). Therefore, the *Regional Occlusion Index* of the ith person is written as

$$o_i = \frac{|O_i|}{|S_i|}, \tag{4.6}$$

where O_i is the set of pixels having positive values in image I_i^O and S_i is the set of pixels having positive values in image I_i^S.

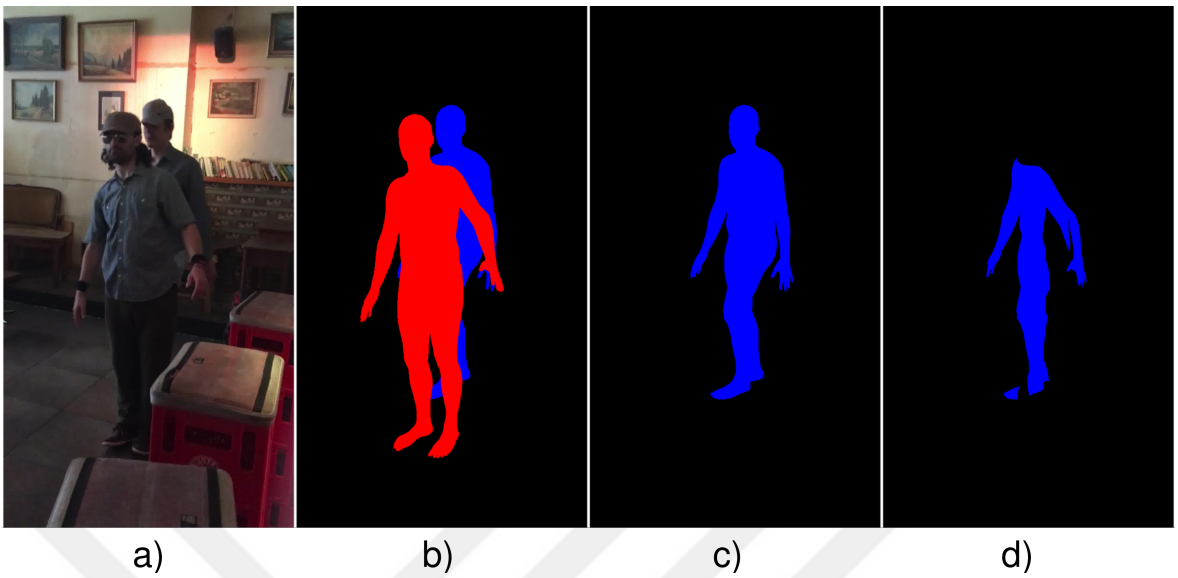

Figure 4.8. Input image (a) render of silhouettes in the input(b), full body render of the occluded person(c), occluded only region of occluded person(d).

4.1.3. Body Part Occlusion Index

The inherent flexibility of the human body presents a challenge to the commonly used skeletal representation, which has limitations in accurately representing different poses. Different body segments have different ranges of motion and articulation capabilities. Specific body segments provide more information about hidden parts, with highly flexible areas conveying critical details compared to less articulated segments. For example, predicting the position of the torso becomes easier when both arms are visible, while observing only the torso makes it challenging to determine the configuration of hidden arms. In Figure 4.9, nearly half of the child's pose at the back is occluded, but the pose is trivial since the arms and legs are visible. Therefore, it is critical to not only assess the degree of occlusion but also to classify the specific type of occlusion.

Although the Occlusion Index (Section 4.1.1) and Regional Occlusion Index (Section 4.1.2) outperform the existing Crowd Index (Section 3.3) in many cases, they overlook the nuanced distinction of the amount of information carried by different body segments and collapse all information into a single score. However, the sensitivity of

the 3D HPS estimators to the visibility of different body regions may differ because of different degrees of articulation. The human body, by nature, does not have the capability of uniform articulation. While some of the joints can rotate around a unit circle, most of the joint's rotation space is limited.

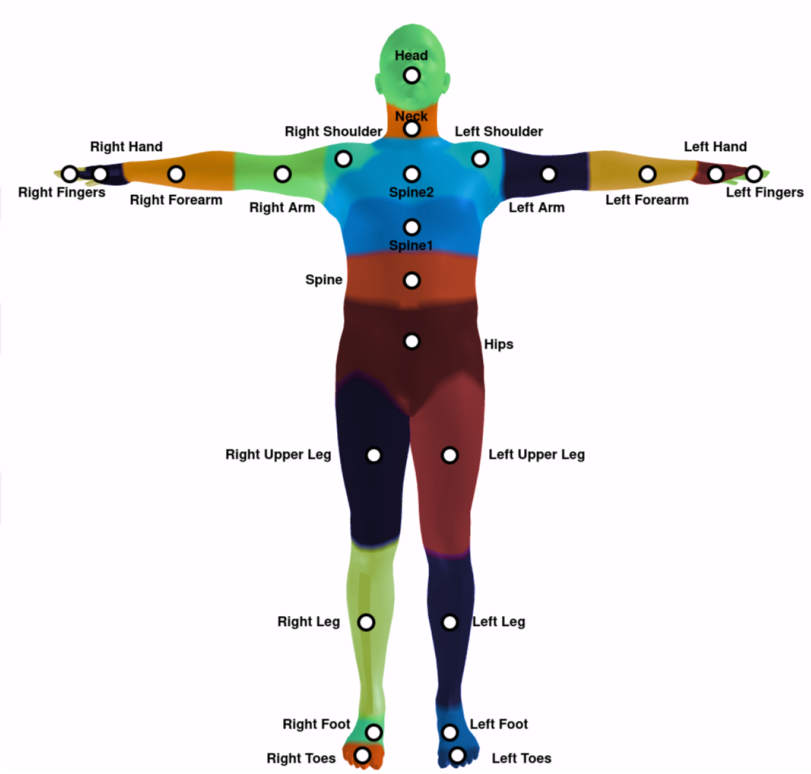

Figure 4.9. Different body segments carry different amounts of information about the pose.

To emphasize this differentiation, we introduce the *Body Part Occlusion Index* (BPOI) to quantify the occlusion for different body parts. This approach allows human datasets to be segmented according to the specific occluded body regions. The methodology thoroughly examines each joint within each person present in the scene. It assesses whether the pixel coordinates of a joint intersect the instance segmentation mask associated with the individual to which the joint belongs. If the joint coordinates lie within the boundaries of the corresponding mask, they are classified as visible; otherwise, they are classified as occluded. This procedure creates an occlusion filter for each person in the scene and eases the categorization.

The Body Part Occlusion Index also allows you to quantify truncation. Truncation is the situation where a person (or object) partially falls outside of the canvas. It can be interpreted as a form of occlusion. BPOI quantifies truncation for each person and joint. This is possible because 2D pose keypoints are projected from 3D. Since the 3D pose is provided regardless of the position of the people in the datasets, their projection is possible even if they are out of frame. Therefore, each joint is classified as either visible, occluded, or truncated. In the Occlusion Index experiments, we assumed that truncated joints were also occluded. To investigate the effect of visibility of different body regions, we manually grouped joints based on their relationship and relative position. Our handcrafted groups are *Upper Body*, *Lower Body*, *Left Body*, *Right Body*, and *Torso*. Figure 4.10 shows these groups visually.

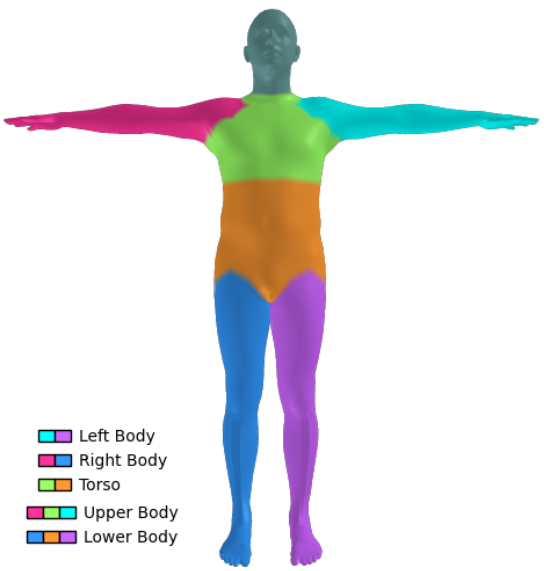

Figure 4.10. Handcrafted joint groups of body part occlusion index.

The groups affect the way we calculate the occlusion score for a person. The vanilla version of the Occlusion Index treats the occlusion of each joint equally. This means that if 12 of the 24 joints are occluded, then the Occlusion Score is 50%. However, the Body Part Occlusion Index only evaluates the joints in the specified group. For example, when forming subsets according to *Torso Group*, it is treated as if the entire joint list consisted of *Spine1*, *Spine2*, *Spine3*, *Pelvis*, *Left Collar*, *Right Collar*, *Left Shoulder*, *Right Shoulder*, *Left Hip*, and *Right Hip* joints. Therefore, in such a case,

if a person's *Right Shoulder*, *Right Elbow*, and *Right Wrist* are occluded, that person's Occlusion score is 10% because only the *Right Shoulder* is present and occluded at the same time in the *Torso Group*. In other words, in the Occlusion Index we take the equally weighted sum of all joints, whereas in the Body Part Occlusion Index we take the equally weighted sum of only the group. This interpretation of the weighting opens a strong path which we will discuss in the next sections.

Note that the same equation (Equation 4.3) also applies to this approach. The only addition is that the joint set is restricted in order to filter which joint is occluded or not. Therefore, no additional processing power is consumed. The formula is

$$k_i^q = \{k_i^j | k_i^j \notin M_i \wedge k_i^j \in K_i^q\},\ j = 1, ..., m^q, \tag{4.7}$$

where m^q is the total number of joints of the group q, K_i^q is the set of joints within group q, meaning $|K_i^q| = m^q$, k_i^j is the jth joint, and k_i is the subset of joints that do *not* fall on the ith person's mask M_i. The joint positions are shown in the Figure 2.1.

The authors of the SMPL provided a map between each vertex on the 3D human mesh model and a joint name. Figure 4.11 visually demonstrates the mapping. Using this mapping, we rendered each body segment separately and calculated the occlusion for each body segment in the same way as described in Section 4.1.2.

The occlusion score o_i^j for person i's jth joint is

$$o_i^j = \frac{|O_i^j|}{|S_i^j|}, \tag{4.8}$$

where S_i^j is the set of pixels whose color is the same as the silhouette, O_i^j is the set of pixels that belong to the occluded region for joint j of person i and c^j is the color for that body segment. The set of pixels from the joint j on the silhouette image is

$$S_i^j = \{(x, y) | I_i^S(x, y) = c^j\}. \tag{4.9}$$

Similarly, the set of pixels representing the occluded body segment j on the occluded region image can be expressed as

$$O_i^j = \{(x, y) | I_i^O(x, y) = c^j\}. \tag{4.10}$$

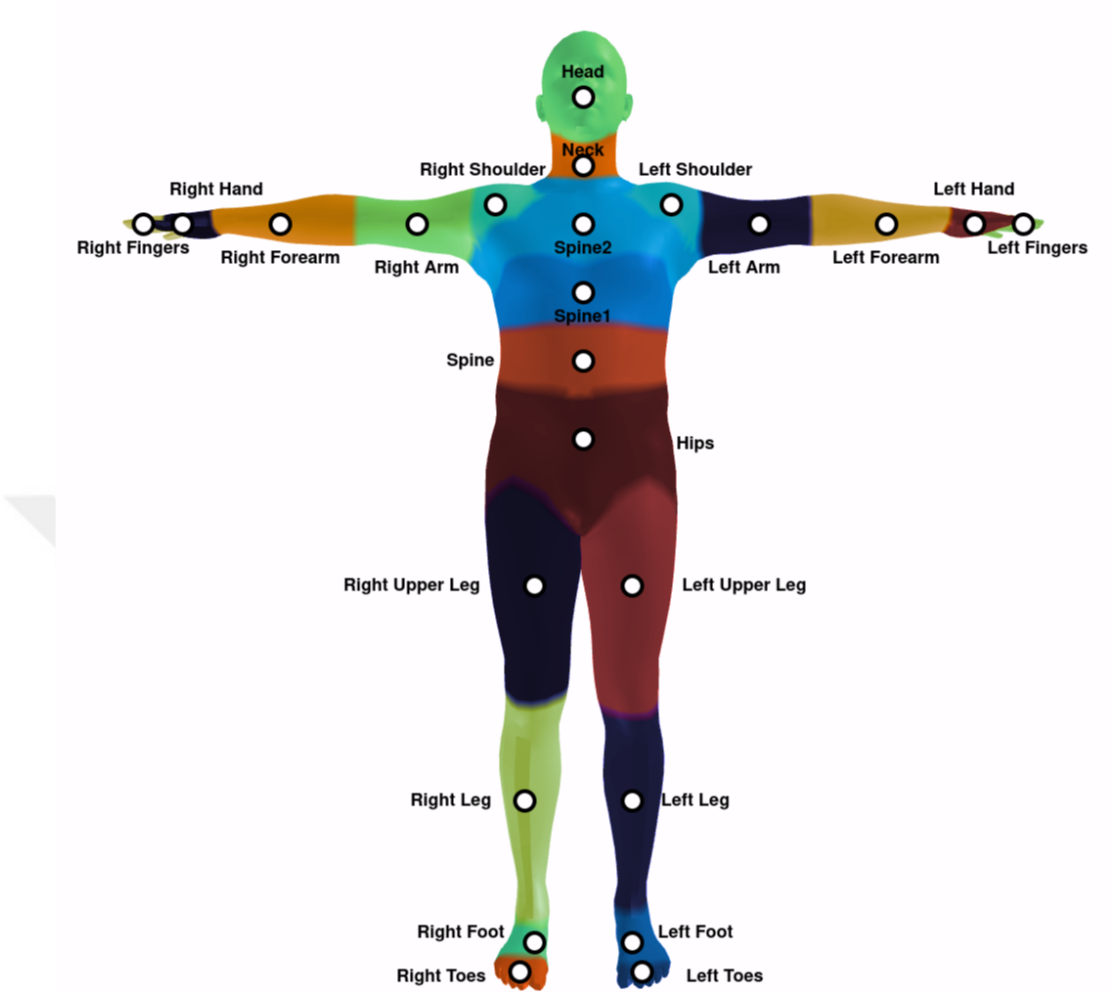

Figure 4.11. The mapping of vertices to body segments.

Figure 4.12 shows an example of rendering each body segment with a different color. Figure 4.12a shows the input image, 4.12b shows the rendered pseudo-instance masks for both people, 4.12c shows the rendering of each body segment with a different color for the person at the back, and 4.12d shows the occluded parts. For example, the person's *Right Forearm* is completely occluded. Therefore, it is fully drawn in 4.12d. Therefore, our index produces a 100% occlusion score for this person's right forearm. On the other hand, *Left Forearm* is not fully occluded, so the ratio of the number of pixels belonging to the left forearm in 4.12d to the number of pixels belonging to the left forearm in 4.12c is calculated as 56% by Regional Occlusion Index.

To distinguish between the rendering of different body parts, the labels for vertices are converted to labels for mesh triangles. Each triangle is classified by majority voting.

If at least two of the three corners of the triangle belong to the same class, the class of that face is fixed. Otherwise, the class is chosen randomly from one of the candidate classes. This feature results in a jagged rendering of the transition zones between two adjacent body segments (See Figure 4.12).

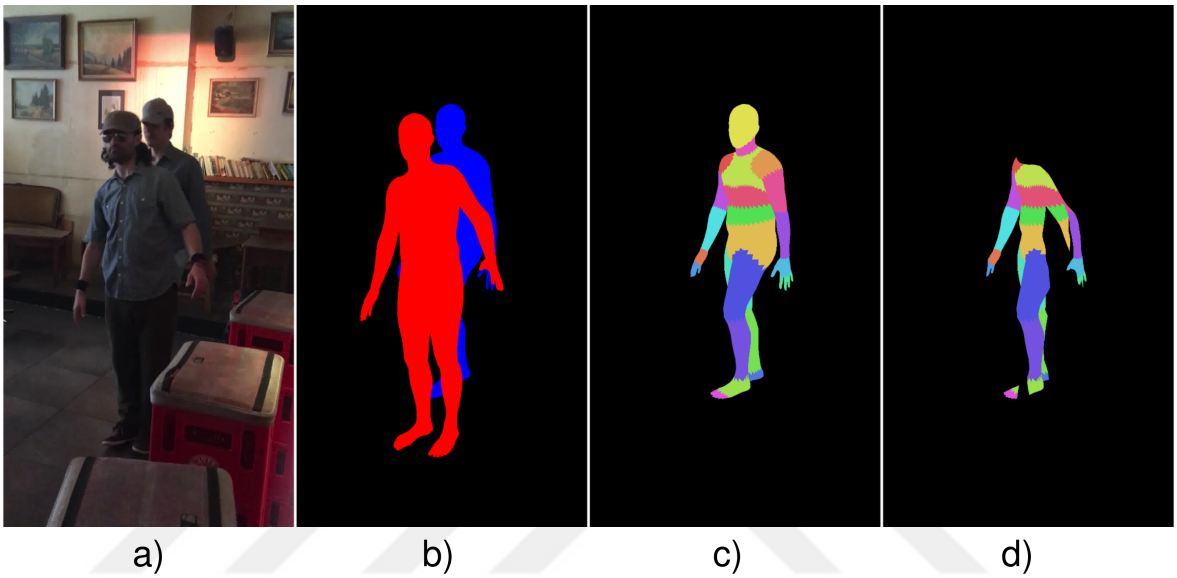

Figure 4.12. Input image(a), silhouettes of both(b), body segments of the occluded(c), occluded only silhouette of occluded(d).

The score of the group with the Regional Occlusion Index is calculated as

$$d^{jq} = \begin{cases} 1 & \text{If joint j is in group q} \\ 0 & \text{If joint j is not in group q} \end{cases}, \tag{4.11}$$

where d^{jq} is the indicator function determining whether joint j is in the group q. The indicator function d^{jq} provides a binary filtering over occlusion scores. For instance, for *Right Body* group, the function returns 1 for *Right Leg* but returns 0 for *Left Leg*. Therefore, the function combining the occlusion scores of body segments for a group is

$$o_i^q = \frac{\sum_{j=1}^{m} d^{jq} * o_i^j}{m^q}, \tag{4.12}$$

where m^q is the number of joints in group q, o_i^j is the occlusion score of joint j of person i, and o_i^q is the occlusion score of person i with respect to group q. Basically, this function calculates the equally weighted average of the selected joints' occlusion scores using the indicator function, occlusion score, and number of joints in the group.

4.1.4. Weighted Occlusion Index

The BPOI calculates just an equally weighted sum of occlusion within a group using the Equation 4.12. The indicator function d can be thought of as a function taking two arguments, being joint and group IDs, and returning either zero or one depending on that joint included in the group. It can be described as

$$d^{jq} : \{(j, q) \mid j \in \{1, ..., m\}, q \in \{1, ...5\}\} \rightarrow \{0, 1\}, \tag{4.13}$$

where j is the joint id and q is the group id.

However, the binary mask nature of the hand-crafted groups in Table **??** oversimplifies the problem and forces us to estimate important joints and link joints as groups based on their spatial location. Instead, an alternative function d, producing a real number weight that scales each body segment's occlusion, can be defined as

$$d^{j} : \{j \mid j \in \{1, ..., m\}\} \rightarrow \mathbf{R}^{[0,1]}. \tag{4.14}$$

In such case, the equation (Equation 4.12) turns into

$$o_i = \frac{\sum_{j=1}^{m} d^j * o_i^j}{m}, \tag{4.15}$$

where m is the total number of joints (24 in the case of SMPL pose format), d^j is the function that produces the weight coefficient for joint j, and o_i^j is the occlusion score for joint j of person i.

The goal is to find a real number coefficient for each joint that will be used as the weight of the occlusion score for that joint. To do this, we exploited the occlusion scores per body segment of the samples having the highest error. This can be achieved by taking the most challenging top K samples and the mean of the occlusion scores per joint previously calculated for those samples. These average occlusion scores per joint are considered the new-defined group's weight.

Figure 4.13 summarizes the joint weight learning process. The first thing to do is to determine the hyperparameter K representing the size of the samples from which

the weights will be learned. Then, the most challenging top K samples are taken according to a 3D HPS estimation method. We take the joint-wise Regional Occlusion Score (o_i^j) for each sample ($\in \mathbf{R}^{Kx24}$) and take the average ($\in \mathbf{R}^{1x24}$). We treat these average occlusion scores as joint weights and rank the samples based on these weights. However, one must be careful when choosing K to avoid the pitfalls. If a high value is chosen, which may be close to the number of samples in the data set, the resulting list will be the same as the list by error. Therefore, K should be as small as possible. Another pitfall is to use the same 3D HPS method for determining weights and weight validation. To ensure the transferability of the learned weights, the method that gives the weights and the method that tests those weights must be different.

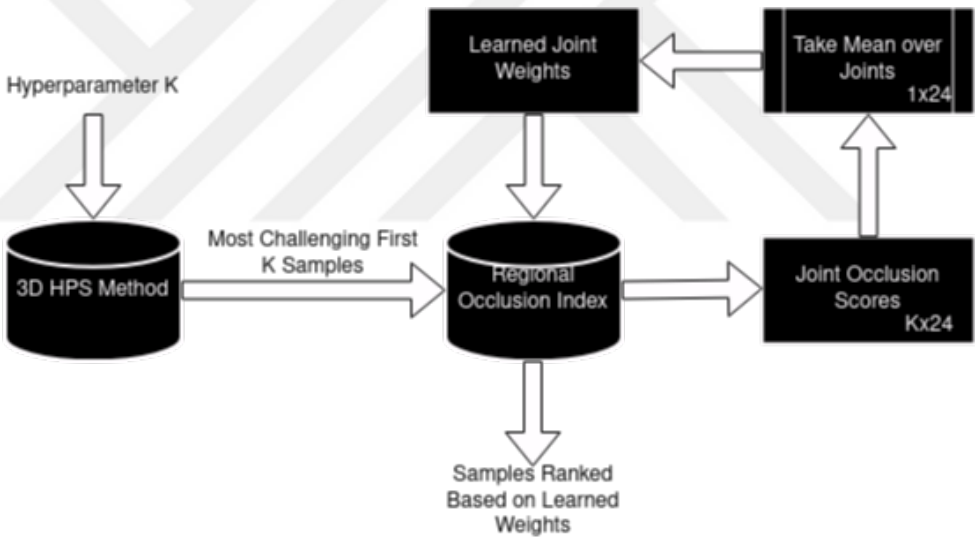

Figure 4.13. The framework of learning joint weights.

4.2. Self-Occlusion Index

Single-view 3D reconstruction is, by definition, an ill-posed problem. There is always an invisible side of a person, no matter which view is taken. Therefore, we extend our index to the case of self-occlusion to further investigate the brittleness of 3D HPS estimators. In this section, we investigate the self-occlusion problem in the case of 3D HPS and quantify the self-occlusion using the index we proposed. We also stressed the effect of root body orientation on the accuracy of the estimators.

4.2.1. Body Orientation Clustering

To investigate the impact of self-occlusion based on body pose, we examined the effect of body orientation on the accuracy of 3D HPS. The process of reconstructing the 3D human body from a single image is inherently challenging and pose sensitive. This challenge arises because individuals inevitably have regions that are self-occluded, where only the side facing the camera is visible. Especially in scenarios where highly flexible body regions are occluded due to the orientation of the body root, accurately estimating the positions of these occluded regions becomes a complex task without contextual information. This approach underscores the importance of considering body orientation in 3D HPS reconstruction, as it profoundly affects the accuracy and complexity of the task.

We collected each person's orientation on the Z-axis (yaw angle) as data points and applied K-means [73] clustering to this data. This clustering allowed us to divide the people in the dataset based on their orientation concerning the camera coordinate frame. Therefore, each person in the dataset is labeled based on the angle between their orientation and the image plane.

The number of clusters is manually adjusted and visually verified. A total of seven clusters are formed, and it is observed that the cluster centers are almost uniformly placed around a full circle, like a person turning around. Figure 4.14 shows the SMPL models representing the cluster centers with their corresponding yaw angle in degrees. Note that the human model facing the camera represents the angle around zero degrees, like the middle one.

Figure 4.14. Cluster centers of 3DPW dataset based on body orientation.

4.2.2. Self-Occlusion Index Per Body Segment

The pipeline described for the Regional Occlusion Index (Section 4.1.2) assigns a class to each triangle of the 3D human mesh, even if another part of the same person occludes it. We collect the statistics of the pixels of such triangles to measure self-occlusion. In this case, however, the triangles at the back of the human surface are always occluded by the triangles at the front. To avoid inflated and misleading statistics, we apply the following filtering. When a pixel is replaced during Z-buffer rendering, we compare the current class and the replacing class. If they are the same, it is not counted as self-occluded. If they are not the same class, it means that a pixel of one body segment is occluded by another body segment of the same person. Therefore, it is counted as self-occlusion.

One problem is the jagged transition zones. Since the rendering projection is done using a perspective camera, almost all of the back surfaces of a body segment are partially occluded by the front surface of a neighboring segment. This leads to an increase in the self-occlusion value for that region. Therefore, we applied a threshold to self-occlusion and did not consider those with a value less than 40% to be self-occluded. This threshold is determined manually and verified visually.

Figure 4.15. Self-occlusion per body segment explained.

Figure 4.15 shows an example body segment for the case of self-occlusion. While the occlusion score for each body segment is zero, the self-occlusion score for *Right Arm* is 100%. Note that the index produces a 14% self-occlusion score for *Left Arm*

because the transitions between adjacent body segments lead to minimal self-occlusion due to perspective projection. The threshold described above encounters such noise.

4.3. A Better Evaluation Metric: Modified MPJPE

Single-view reconstruction of humans under occlusion is usually not a trivial task. 3D HPS estimation models are expected to estimate the invisible part. This might be fair for some cases where the occlusion is minimal, or the occluded part is just the middle of a straight bone, so there is not much variance for that region. However, this may not be the case for occlusions of highly articulated body segments, where the models often fail, and we are more interested.

In the case of severe occlusion, the expectation becomes an estimate of what the person is doing in that frame. For example, if a person's legs are not visible, they could be sitting, walking, or standing still. Since the goal is single-view reconstruction, models cannot interpolate between the previous or next frame with two visible poses. Using contextual information can shed light on the process. For example, if the frame in a basketball game is captured when a person's arms are partially occluded, we can assume that the person is in the pose of holding a ball. However, explicitly preserving context is not yet common practice for 3D HPS. Moreover, for some datasets, the pose of the occluded region is also unknown to the annotator, as well. For those cases, the 3D HPS problem partially turns into approximating what the annotator thought while annotating the data, which might be erroneous, too.

Therefore, evaluating models based on their performance on the visible joints but not the occluded ones would be a fairer approach, and our Occlusion Index allows us to do so. The expectancy from 3D HPS methods should not be guessing what the annotator thought but reconstructing the visible body segments regardless of the presence of occlusion. Therefore, we propose to change the definition of occlusion robustness slightly. The current definition can be summarized as "The ability of a 3D HPS estimator to estimate the properties of occluded regions as good as visible

ones." Instead, we offer " *The ability of a 3D HPS estimator to estimate the properties of visible regions regardless of the presence of occlusion in other regions.*"

The commonly used metric for human pose evaluation is the *Mean Per Joint Position Error (MPJPE)*. The error is the Euclidean Distance (L2) between the predicted and the groundtruth joints. It can be formalized as

$$\text{MPJPE}=\frac{1}{m}\sum_{j=1}^{m}\sqrt{\sum_{q=1}^{3}(\mathbf{v}^{jq}-\mathbf{w}^{jq})^2}, \tag{4.16}$$

where j is joint id, q is the dimension id and $\mathbf{v^{jq}}$ and $\mathbf{w^{jq}}$ are the scalars for prediction and groundtruth for jth joint's qth dimension, respectively.

What we offer instead is the *Modified MPJPE*

$$\text{Modified MPJPE}=\frac{1}{m}\sum_{j=1}^{m}(1-o^j)*\sqrt{\sum_{q=1}^{3}(\mathbf{v}^{jq}-\mathbf{w}^{jq})^2}. \tag{4.17}$$

Note that while o^j represents the occlusion score for joint j (either Occlusion Score or Regional Occlusion Score), $(1-o^j)$ represents the corresponding joint's visibility. Remember that our approach makes this modification possible.

5. EXPERIMENTS

To the best of our knowledge, there is currently no proposed experimental setup for comparing the precision of indices that quantify occlusion in the 3D Human Pose and Shape Estimation datasets. Therefore, we propose a novel experimental setup to enable the comparison of such indices. This approach is also suitable for comparing indices in other scenarios where the goal is to sample the most challenging or easiest samples.

The quantification of occlusion for each individual in the image is prioritized over generating a full score for the image. This approach ensures that the core idea of distinguishing between occluded and occluding persons is not violated. For instance, if there are two people in the image and one completely occludes the other, the average occlusion for the image would be 0.5 (assuming the occlusion index is normalized between 0 and 1), which might not be considered challenging. However, calculating the occlusion index for each individual in the image is crucial, especially when dealing with fully occluded individuals. Averaging the entire image should be avoided.

The amount of occlusion ranks the samples in the dataset in descending order. They are then fed into the 3D HPS model one at a time, starting at the top. The error is expected to start at the highest and slowly decrease as the occlusion value decreases through the list. The change in error value between two different occlusion quantifiers is compared, with the more accurate one selecting the more difficult samples at the beginning of the process. Both algorithms are expected to converge to the same average error as they process the entire data set. However, the stronger candidate will likely cause more errors in the early stages.

We report the running mean error to emphasize the cumulative difference between the two subsets. Our goal is not to find the most occluded single sample but rather to form a benchmark list of samples large enough to collect statistics. We report

the average error for the dataset's most occluded top K percent, where we expect the stronger candidate to cause more average error when K is low. We also visually examine the error with graphs.

We conducted our experiments on 3DPW [30], AGORA [29], and OCHuman [31] datasets. Unless otherwise specified, we utilized courtyard_warmWelcome_00, courtyard_dancing_01, courtyard_basketball_00, downtown_bar_00, courtyard_goodNews_00, courtyard_captureSelfies_00, courtyard_giveDirections_00, courtyard_hug_00, courtyard _dancing_00, and courtyard_shakeHands_00 scenes in 3DPW; archviz and hdri_50mm scenes of AGORA; and the samples whose SMPL parameters are successfully provided by EFT of OCHuman. For this study, we excluded scenes that either had no occlusion or had too many people in the background without providing annotations. It is important to note that the number of reliable SMPL annotations for OCHuman was quite low, so we only used OCHuman in the first experimental setup.

5.1. Details of The Experiments

5.1.1. Procrustes Aligned MPJPE

It is a common practice to transform the prediction to be aligned with the groundtruth before calculating the error. The reason for this is to obtain robustness to scale and translation. The most popular technique for aligning two 3D models is to use Procrustes Alignment (PA) [74]. The Procrustes procedure rotates and scales a given prediction to be similar to the given ground truth. Figure 5.1 shows an example result of PA. 5.1a (red) represents a groundtruth 3D model and 5.1b (blue) represents a prediction that is the 45 degree rotated version of the groundtruth around the z-axis. 5.1c (gray) is the PA version of the prediction where it is perfectly fitted to the groundtruth. Note that the MPJPE for this original prediction (b) is 0.75 and the MPJPE for the PA version (c) is 0. The MPJPE reported for the PA version is often referred to as PA-MPJPE. It is generally preferred over the original MPJPE since the SMPL format represents the relative relation between neighboring joints.

Figure 5.1. Ground-truth(a), 45 degrees rotated(b), and Procrustes aligned(c) poses of the same human model.

5.1.2. Truncation Handling

Truncation can be assumed as a type of occlusion. We proposed an approach to merge truncation and occlusion scores. As a joint cannot be both truncated and occluded in the Occlusion Index, we apply the logical OR between them. However, for the Regional Occlusion Index, both values are real numbers.

The truncation is defined as the ratio between the number of pixels that fall outside the frame (also not visible) and the total number of pixels that a body segment would occupy if the frame were unlimited. On the other hand, occlusion is the ratio between the number of pixels that are not visible due to another person (occluder) and the total number of pixels belonging to that body segment within the frame (would be visible if there was no occluder). Therefore, to merge the truncation and occlusion statistics, we scaled the occlusion by the non-truncated part as

$$O^j = [(1 - t^j) * o^j] + t^j, \tag{5.1}$$

where o^j and t^j are the occlusion scores and truncation of joint j, respectively. Consider a case where a body segment is 50% truncated, and 50% of the in-frame region is occluded. This means only the 25% of the segment is visible and 75% or the segment is not visible due to different reasons. Equation 5.1 gives a consistent result $O^j = [(1 - 0.5) * 0.5] + 0.5 = 0.75$.

5.1.3. Matching Procedure

The 3D annotation of people in the image is given in an order in terms of data structure. However, the order or prediction for these people does not necessarily match the ground truth order. Therefore, a matching mechanism between the predicted parameters and the ground truth parameters is needed. In addition, the 3D HPS models may not find all the people in the image or may find more 3D models than they should. As a result, a matching mechanism between the predictions and the groundtruths should be established. Let the prediction tensor be $P \in \mathbb{R}^{Nx24x3}$ and the ground truth tensor be $G \in \mathbb{R}^{Mx24x3}$. We keep the groundtruth order fixed and match each groundtruth to a prediction based on the distance between 2D keypoints. If there are too many predictions ($N > M$), we match those predictions to the closest groundtruth. If there is not enough number of predictions ($N < M$), it means that a person in the image could not be found, which often happens under occlusion; we match this groundtruth to the neutral pose, where all SMPL pose parameters are zero. Note that for the people in the image not found by the 3D HPS model, we set 5.0 MPJPE and 1.0 PA-MPJPE.

5.2. Performance on Entire Body

In this section, we report on the performance of our indices and the existing Crowd Index on different datasets. In these experiments, body regions are not distinguished, and these indices generate a single occlusion value per person.

Each index (OI, ROI, CI) creates its own ranked list from the data set. We start to evaluate images from the top of this list. In the tables shown in this section, the top $K\%$ samples are the subset of the data set when ranked by different indices based on the amount of occlusion. The top K% of the dataset does not mean the same samples for each index, but each of them contains an equal number of examples. For example, for the cell at the intersection of K is 15, the error type is MPJPE, the method is ROMP, and the index is OI, we sort each sample in the data set (remember, by sample, we

mean the people, not the frames) based on the score generated by the Occlusion Index, we take the first 15% of this ranked list. Then, we take the corresponding error that ROMP made for each sample, and the value in the cell shows the average of those error values. Therefore, the higher the score, the more difficult the subset.

5.2.1. OCHuman

OCHuman has a slightly different pose format than other datasets. OCHuman provides 2D poses in COCO [45] format ($\mathbb{R}^{19x2}$) (see Figure 5.2a). However, 3D HPS methods provide poses in SMPL format ($\mathbb{R}^{24x2}$), as shown in Figure 2.1. Therefore, to be able to use the matching algorithm, we need to convert one to the other.

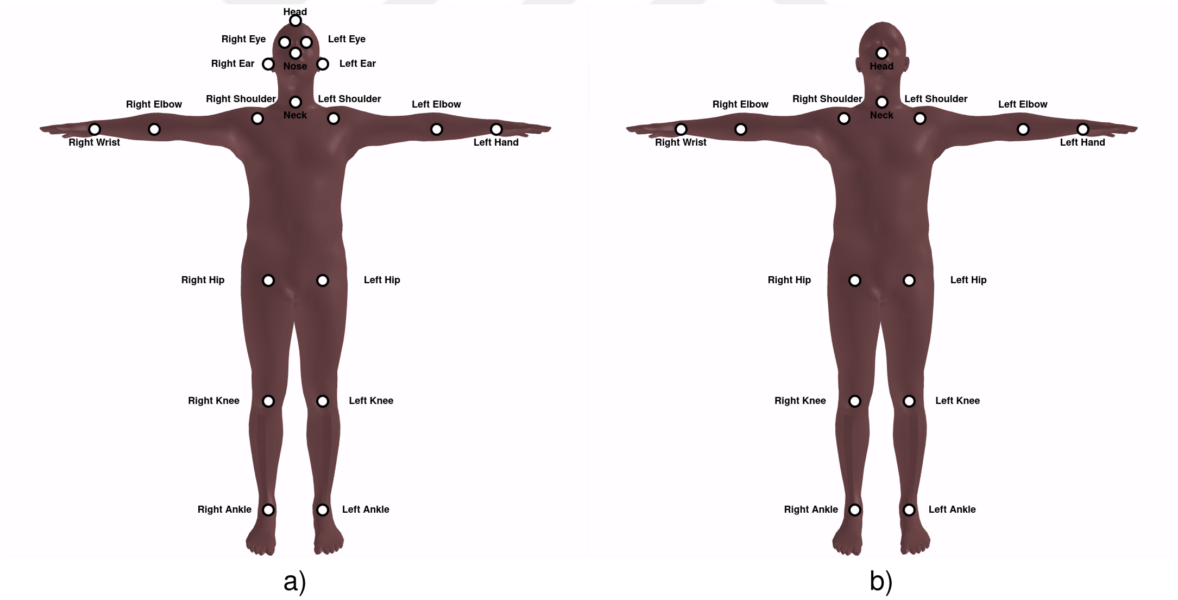

Figure 5.2. The original COCO pose format(a) and the modified version(b).

We chose to convert the SMPL pose to the COCO format because the COCO format does not have a joint near the spine and collar. However, the SMPL format does not include some of the COCO format's key points on the face, such as eyes and ears. Nevertheless, we believe that their occlusion or visibility can be simplified by the head. Therefore, we trim the COCO format to 14 joints by excluding *Right Eye*, *Left Eye*, *Right Ear*, *Left Ear*, and *Head*. (See Figure 5.2b) Note that we take the

Nose keypoint of COCO as the *Head* of SMPL format since it is placed at the center of the face similar to the *Head* keypoint of SMPL, while the *Head* keypoint of COCO represents the top of the head, which is further away from the *Head* keypoint of the SMPL pose. We take the remaining first 14 joints of the COCO pose format directly since all of them have an SMPL equivalent. Finally, there are some missing keypoints in the original annotation of OCHuman. These keypoints are marked as $(0, 0)$, and we set the corresponding modified COCO joint parameters in the 14 joints to $(0, 0)$ as well, to accurately match predictions and groundtruths.

Table 5.1. Running mean error on OCHuman dataset for entire body.

			Top K Percent						
Error Type	**Method**	**Index**	**10**	**15**	**20**	**25**	**30**	**40**	**50**
MPJPE	**ROMP**	**OI**	**2.63**	2.61	2.53	2.39	2.41	2.45	2.32
		CI	2.25	2.36	2.35	2.28	2.23	2.28	2.29
	BEV	**OI**	2.87	**2.94**	2.83	2.63	2.61	2.63	2.55
		CI	2.84	2.72	2.73	2.61	2.57	2.59	2.66
PA-MPJPE	**ROMP**	**OI**	**0.41**	0.40	0.40	0.38	0.37	0.37	0.35
		CI	0.35	0.36	0.36	0.35	0.35	0.35	0.35
	BEV	**OI**	**0.41**	0.41	0.41	0.40	0.39	0.39	0.38
		CI	0.40	0.40	0.40	0.39	0.39	0.39	0.39

We took samples of the OCHuman dataset if their SMPL annotations provided by EFT [25] are valid. We check the validity by comparing the number of models indicated by the 2D pose annotation of the original dataset with the number of models indicated by the SMPL parameters. If they match, we include that sample in our validated subset. We also excluded some of the samples due to incorrect instance mask annotation. We detected the existence of such samples by visually examining the samples where our Occlusion Index produces a high value, but the methods do not suffer from performance degradation, and vice versa. The list of these samples and visual examples are given in the Appendix A. As a result, we ran OCHuman experiments on 1047 samples in 546 frames.

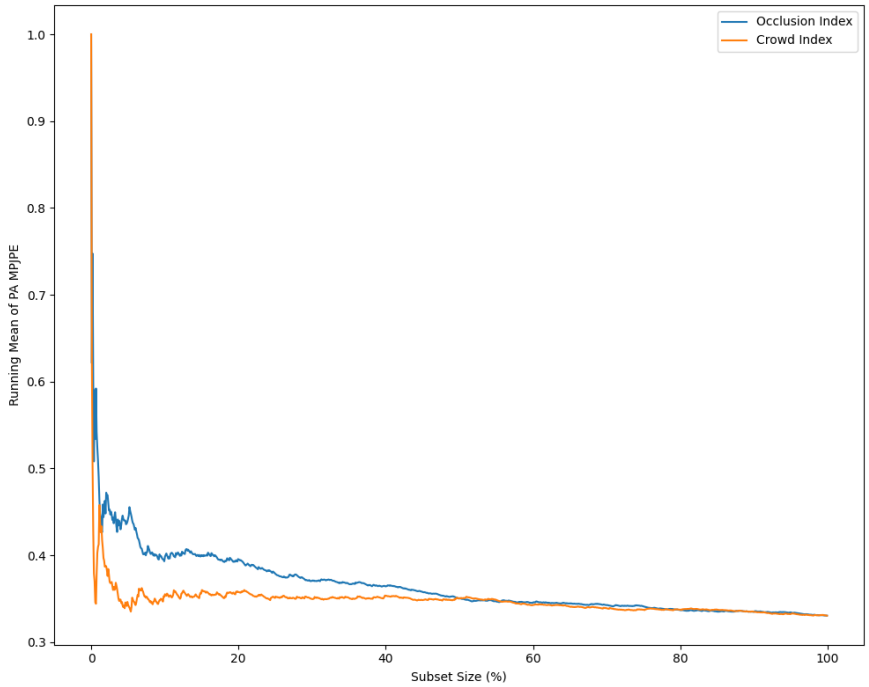

Figure 5.3. ROMP's PA-MPJPE on OCHuman.

Table 5.1 shows the performance of ROMP on samples ranked by our Occlusion Index (OI) and Crowd Index (CI). Our Occlusion Index is able to select more challenging samples across all error types and methods. Therefore, the subsets collected by the Occlusion Index led methods to fail more. Figure 5.3 shows the running mean PA-MPJPE of ROMP on the different subsets formed by our Occlusion Index and the Crowd Index. The error OI led is higher than the error CI caused, especially at the beginning of the ranked lists, corresponding to the samples with the most occlusion scores. We did not apply the Regional Occlusion Index (Section 4.1.2) to OCHuman because the camera parameters provided by EFT are not consistent enough to be projected onto the same frame. Although the way the dataset is collected is quite suitable for the case of occlusion since this dataset is marker-less, the reliability of the 3D annotation is questionable.

5.2.2. AGORA

AGORA provides all the necessary annotations for this study. They provide 2D and 3D keypoints in SMPL-X [43] format, where the first 24 joints are equal to SMPL

format, instance segmentation masks, and SMPL parameters for each person. Besides, since it is a simulation dataset, the annotation does not suffer from noisy annotations. We use the same experimental setup described at the beginning of this chapter. We rank the dataset using the Crowd Index, Occlusion Index, and Regional Occlusion Index, and compare the performance of the method on each sample by aggregating them into a running mean.

Table 5.2. Running mean error on AGORA dataset for entire body.

			Top K Percent						
Error Type	**Method**	**Index**	**10**	**20**	**30**	**40**	**50**	**60**	**70**
MPJPE	**ROMP**	**ROI**	**4.83**	4.71	4.46	4.22	4.01	3.86	3.77
		OI	4.82	4.61	4.32	4.15	4.02	3.92	3.79
		CI	3.75	3.68	3.71	3.72	3.71	3.67	3.64
	BEV	**ROI**	**4.66**	4.43	4.02	3.76	3.60	3.50	3.42
		OI	**4.66**	4.25	3.83	3.65	3.54	3.48	3.40
		CI	3.42	3.34	3.36	3.35	3.33	3.31	3.29
PA-MPJPE	**ROMP**	**ROI**	**0.96**	0.94	0.89	0.83	0.79	0.75	0.73
		OI	**0.96**	0.92	0.86	0.83	0.80	0.77	0.75
		CI	0.73	0.71	0.72	0.72	0.72	0.72	0.71
	BEV	**ROI**	**0.93**	0.88	0.79	0.73	0.69	0.66	0.64
		OI	**0.93**	0.84	0.75	0.71	0.69	0.67	0.65
		CI	0.64	0.63	0.63	0.63	0.63	0.63	0.62

We used AGORA's *archviz* and *hdri_50mm* scenes in the validation set. This makes a total of 3952 people in about 500 frames. Table 5.2 shows the performance of the methods on different subsets. We observe that Occlusion Index and Regional Occlusion Index select the most challenging examples in the first sampling stages, and the error gradually decreases as easier samples are introduced. On the other hand, Crowd Index cannot distinguish the hard and easy samples and performs an average over the dataset to all occlusion levels similar to the random choice. Figure 5.4 visualizes the performance of ROMP with respect to PA-MPJPE on different ranked

subsets by different indices. Note that the Occlusion Index and the Regional Occlusion Index start with the samples with the highest error and gradually converge to the mean error of the dataset. However, the Crowd Index suffers from mediocre sampling and cannot distinguish between a difficult example and an easy one.

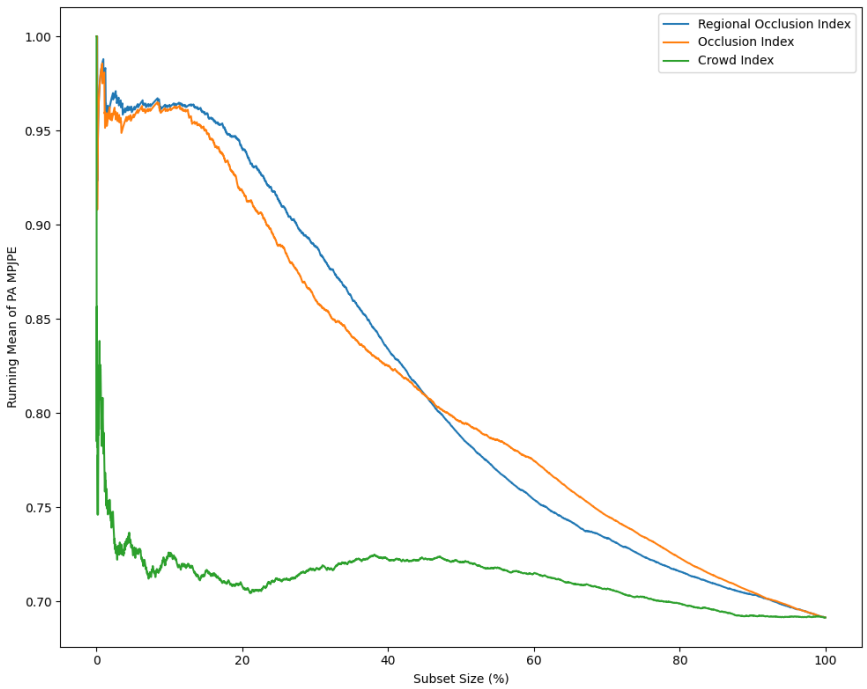

Figure 5.4. ROMP's PA-MPJPE on AGORA.

5.2.3. 3DPW

To compute the Occlusion Index and Regional Occlusion Index on 3DPW, we first took a render of the SMPL bodies in the image plane to be used as pseudo-instance masks. In addition, 3DPW provides the 3D coordinates of the SMPL pose joints in the camera coordinate frame. Therefore, each 3D point representing 24 SMPL joints was projected onto the image plane as 2D pixels. Later, the bounding boxes for the Crowd Index are calculated from the four most distant keypoints of the person with an additional offset.

We used about 12K samples from 6314 images of 10 scenes. We report the performance on ROMP's PA-MPJPE on the first 10% of the ranked lists (about 1.2K

Table 5.3. Running mean error on 3DPW dataset for entire body.

			Top K Percent						
Error Type	**Method**	**Index**	**1**	**1.5**	**2**	**2.5**	**3**	**4**	**5**
MPJPE	**ROMP**	**ROI**	3.62	3.47	3.42	3.34	3.29	3.26	3.24
		OI	**3.66**	3.54	3.44	3.36	3.33	3.28	3.25
		CI	3.33	3.42	3.42	3.43	3.45	3.44	3.38
	BEV	**ROI**	3.65	3.46	3.35	3.29	3.24	3.22	3.22
		OI	**3.70**	3.51	3.40	3.33	3.31	3.26	3.24
		CI	3.31	3.41	3.41	3.42	3.44	3.41	3.35
PA-MPJPE	**ROMP**	**ROI**	**0.51**	0.47	0.46	0.44	0.43	0.42	0.43
		OI	0.50	0.46	0.44	0.43	0.43	0.42	0.42
		CI	0.38	0.38	0.37	0.37	0.37	0.37	0.38
	BEV	**ROI**	**0.53**	0.48	0.45	0.43	0.42	0.42	0.42
		OI	0.52	0.47	0.45	0.43	0.43	0.42	0.41
		CI	0.38	0.38	0.37	0.37	0.37	0.37	0.37

people). We report the first 10% instead of the entire dataset because the indices are already converging to the same level. The reason for the convergence at this stage is the limited number of occluded samples in the entire dataset. Besides, earlier determination of the occluded samples is the main objective. We also excluded the samples with more than 0.75 truncation in 3DPW. 3DPW is a sequence dataset, where the images consist of consecutive frames of a video. Therefore, sometimes the actors leave the scene for some time, but their parameters are provided so that they are technically in the scene. However, it is not realistic to expect methods to find these people, so we used such filtering.

Table 5.3 shows the running average of the errors made by ROMP and BEV. Note that ROI and OI force the methods to make more errors by choosing more difficult samples, especially PA-MPJPE rows with similar scores. On the other hand, CI selects equally challenging samples on all occlusion levels, performing close to random choice. The Crowd Index does worse in this case of 3DPW compared to other datasets. In fact,

the samples that CI reported as the most challenging are simpler than all the samples that other metrics choose. The reason is the *Pseudo Depth Order* principle described in Section 4.1.1. The Crowd Index does not have the ability to distinguish between the occluder and the occluded (see Section 3.3). Therefore, some of the samples it reports are not even occluded so that they can be easily reconstructed.

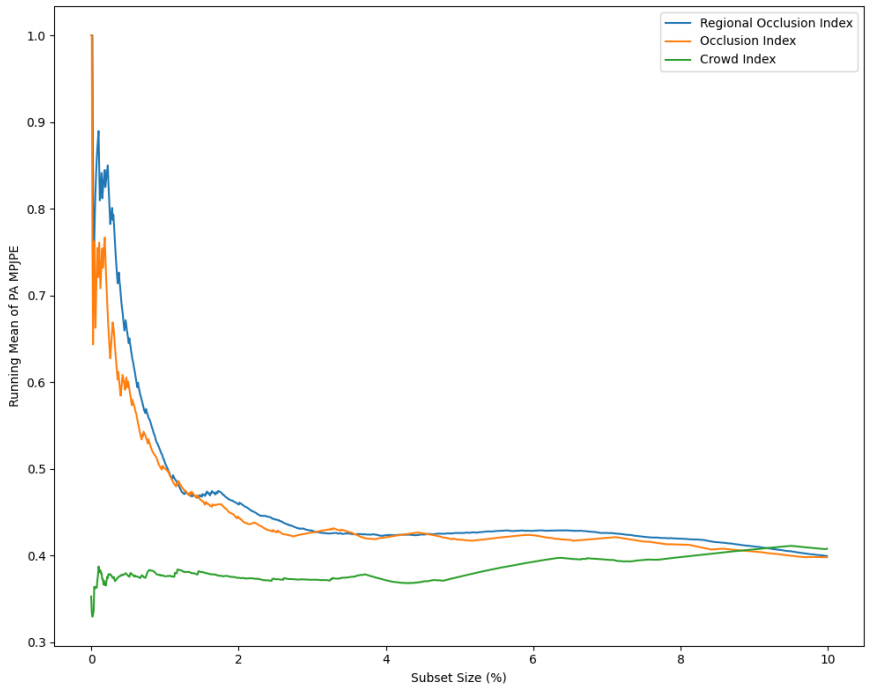

Figure 5.5. ROMP's PA-MPJPE on 3DPW.

Figure 5.5 shows the ROMP's PA-MPJPE visually. We see both the Occlusion Index and Regional Occlusion Index are able to select challenging samples leading to 1.0 PA-MPJPE at the beginning. Remember that if 1.0 MPJPE is produced, it means that the 3D HPS model will not be able to detect this person. Not being able to fully detect a person is a natural consequence of severe occlusion. In this case, we can say that the first examples selected by both the Occlusion Index and the Regional Occlusion Index suffer from extreme occlusion. In fact, we can visually verify this. Figure 5.6 shows one of the most difficult samples according to Crowd Index (a) and Regional Occlusion Index (b). Note that the person with a Crowd Index of 1.00 is the one in front (red) and is fully visible. On the other hand, the Regional Occlusion Index produces an occlusion score of 0.938 for the person at the back (blue) who is severely

occluded. This is a clear example of the superiority of our index due to the *Pseudo Depth Order* principle.

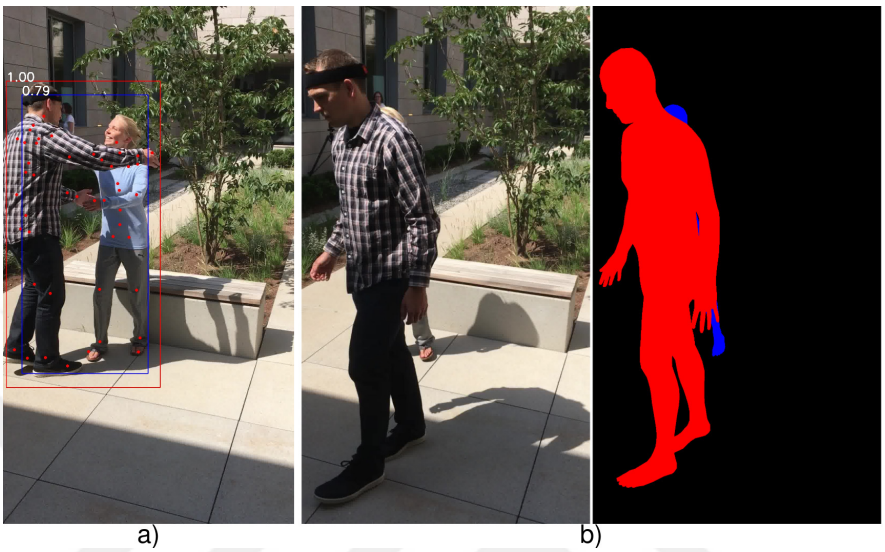

Figure 5.6. The most challenging examples according to Crowd Index(a) and Regional Occlusion Index(b).

In addition to Figure 5.3, we can visually represent the superiority of our metrics in an alternative graph. Figure 5.7 shows the running mean per occlusion score instead of the subset size. The difference between our metrics and the Crowd Index is more noticeable in the case of severe occlusion. Note, however, that the running mean is still based on the most occluded example. For example, if there are N samples represented in the figure, the points closer to the 1.0 occlusion score represent the running mean with fewer samples, and those closer to 0 represent more samples. Therefore, the leftmost sample with the lowest occlusion score represents the average of the entire subset, and the rightmost sample with the highest occlusion score represents only itself. When using the running mean, we need to collapse the errors into a single value at either the lowest occlusion score or the highest occlusion score. We prefer to compute the running mean starting from the highest occlusion and collapse the low occlusion side to one scalar because this version allows us to distinguish samples with high occlusion scores, which is what we are ultimately looking for. Therefore, if the next sample we include increases the average error up to that point, we can say that it

is more challenging for the method. If the average error up to that point decreases after including that sample, then it means that either the method is robust to that type of occlusion or we have misquantified the occlusion. Remember that, other factors, like illumination, can also affect the model's performance.

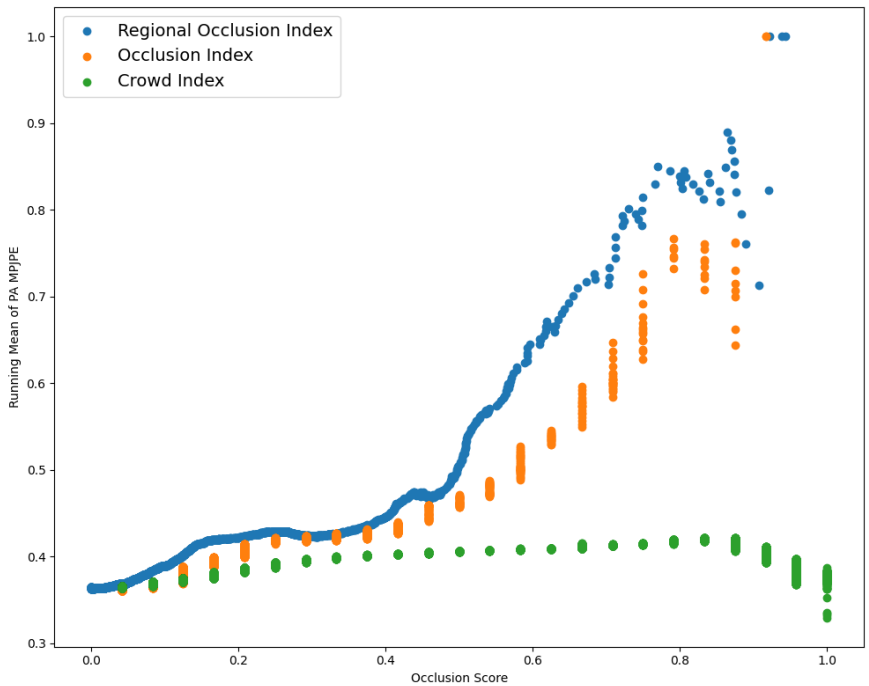

Figure 5.7. ROMP's PA-MPJPE on 3DPW per occlusion score.

Another observation is the discrete and continuous-like scatter of the indices. Since both Crowd Index and Occlusion Index are based on keypoints, it forces the occlusion values to be in certain levels. On the other hand, the Regional Occlusion Index produces a more even distribution of occlusions across the dataset. Note that the occlusion score produced by the Crowd Index does not exceed 1 in this case because the scenes we selected in the 3DPW dataset contain only two people. This is why we used the previous type of graphs for comparison to the Crowd Index.

5.3. Occlusion of Different Body Segments

In Section 4.1.3 the need to distinguish the occlusion of different body parts is explained. To investigate which part of the body carries more information about the

pose and is vital for not being occluded, we manually designed different groups of joints. Note that the joints in each group are spatially related.

Table 5.4. Running mean error on 3DPW dataset for handcrafted groups.

			Occlusion Score						
Error	**Method**	**Group**	**0.2**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
MPJPE	**ROMP**	**Torso**	3.15	3.29	3.39	3.61	3.80	3.97	4.18
		Upper Body	3.16	3.27	3.43	3.82	4.39	4.68	**5.00**
		Lower Body	3.20	3.41	3.61	3.87	4.00	3.72	3.63
		Left Body	3.23	3.29	3.29	3.35	3.29	3.25	3.45
		Right Body	3.13	3.17	3.24	3.28	3.40	3.48	3.68
	BEV	**Torso**	3.15	3.24	3.34	3.54	3.75	4.00	4.44
		Upper Body	3.17	3.30	3.42	3.81	4.43	4.65	**5.00**
		Lower Body	3.17	3.39	3.68	3.86	4.27	4.52	4.35
		Left Body	3.26	3.31	3.33	3.36	3.31	3.30	3.42
		Right Body	3.10	3.13	3.23	3.27	3.40	3.59	4.00
PA-MPJPE	**ROMP**	**Torso**	0.38	0.42	0.45	0.52	0.58	0.63	0.70
		Upper Body	0.39	0.44	0.47	0.59	0.78	0.87	**1.00**
		Lower Body	0.40	0.47	0.52	0.59	0.65	0.56	0.52
		Left Body	0.41	0.44	0.45	0.42	0.38	0.36	0.44
		Right Body	0.36	0.39	0.41	0.42	0.46	0.51	0.57
	BEV	**Torso**	0.38	0.41	0.44	0.51	0.58	0.66	0.80
		Upper Body	0.39	0.44	0.48	0.61	0.80	0.87	**1.00**
		Lower Body	0.40	0.47	0.55	0.60	0.74	0.83	0.77
		Left Body	0.42	0.44	0.45	0.43	0.40	0.39	0.44
		Right Body	0.36	0.39	0.42	0.43	0.47	0.56	0.67

Our proposed experimental setup is the same procedure described at the beginning of this chapter (Chapter 5). The only difference is that we do not care whether the joints that are not in the group are occluded or not. For example, when ranking the samples for the *Torso Group*, we rank the samples based on the occlusion in *Spine1*,

Spine2, *Spine3*, *Pelvis*, *Left Collar*, *Right Collar*, *Left Shoulder*, *Right Shoulder*, *Left Hip*, and *Right Hip*. For example, the occlusion of *Right Foot* or *Left Wrist* does not affect the occlusion score within this group.

Note that this experiment was performed for the 3DPW and AGORA datasets, but not for OCHuman. Due to the limited number of samples with SMPL parameters in OCHuman, the number of samples in all groups is not sufficient to conclude a meaningful statistic. Another remark is we report the occlusion scores that are produced by the Regional Occlusion Index in this section to have continuous plots.

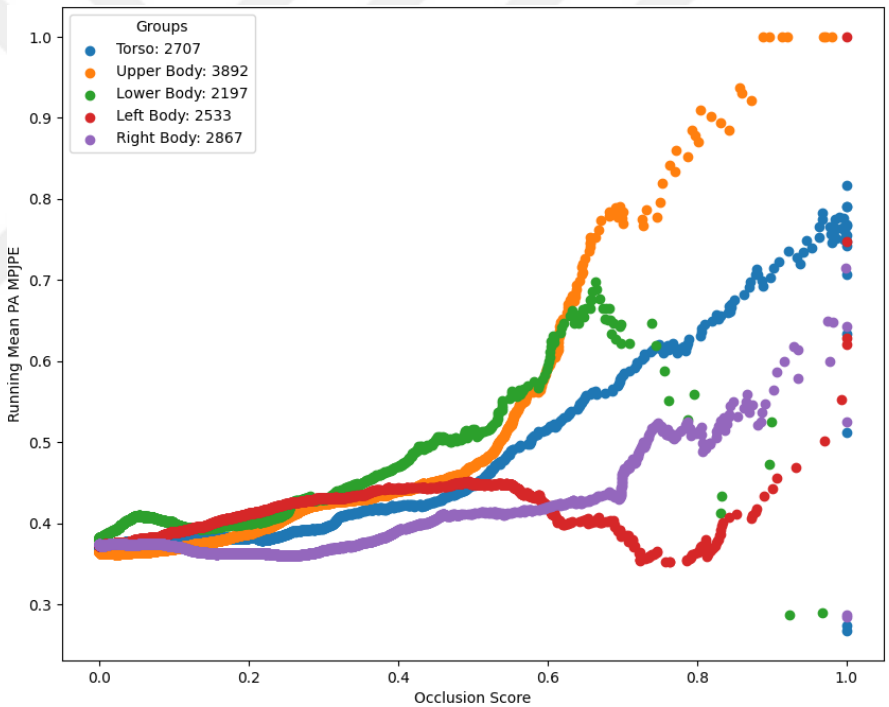

Figure 5.8. ROMP's PA-MPJPE on handcrafted 3DPW groups per occlusion score.

Table 5.4 shows the performance of ROMP and BEV on subsets of 3DPW formed by hand-crafted groups. The results demonstrate that the occlusion of *Upper Body* challenge methods more while *Lower Body* and *Torso* are the second most challenging. However both *Left Body* and *Right Body* can not challenge ROMP especially at high occlusion. We interpret this to mean that there is not a strong bias difference between the left and right sides of the body. Figure 5.8 visually demonstrates the findings of Table 5.4. The numbers in the legend are the number of samples in each subset.

Table 5.5. Running mean error on AGORA dataset for handcrafted groups.

			Occlusion Score						
Error	**Method**	**Group**	**0.2**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
MPJPE	**ROMP**	**Torso**	4.28	4.41	4.47	4.55	4.59	4.66	4.73
		Upper Body	4.24	4.45	4.55	4.63	4.70	4.75	**4.81**
		Lower Body	3.78	3.82	3.85	3.92	4.01	4.14	4.34
		Left Body	3.81	4.10	4.28	4.36	4.46	4.55	4.66
		Right Body	3.82	4.12	4.30	4.39	4.47	4.55	4.64
	BEV	**Torso**	3.78	3.92	4.00	4.09	4.17	4.25	4.35
		Upper Body	3.75	3.96	4.08	4.19	4.30	4.42	**4.54**
		Lower Body	3.46	3.52	3.55	3.62	3.72	3.84	4.05
		Left Body	3.48	3.72	3.88	3.97	4.11	4.22	4.37
		Right Body	3.48	3.71	3.86	3.97	4.10	4.23	4.39
PA-MPJPE	**ROMP**	**Torso**	0.85	0.88	0.89	0.91	0.92	0.93	0.94
		Upper Body	0.84	0.89	0.91	0.92	0.94	0.95	**0.96**
		Lower Body	0.74	0.74	0.75	0.77	0.79	0.82	0.86
		Left Body	0.74	0.81	0.85	0.86	0.89	0.91	0.93
		Right Body	0.74	0.81	0.85	0.87	0.89	0.90	0.92
	BEV	**Torso**	0.74	0.77	0.79	0.81	0.82	0.84	0.86
		Upper Body	0.73	0.78	0.80	0.83	0.85	0.88	**0.90**
		Lower Body	0.65	0.66	0.67	0.69	0.71	0.74	0.79
		Left Body	0.66	0.72	0.75	0.77	0.81	0.83	0.86
		Right Body	0.66	0.72	0.75	0.78	0.81	0.83	0.87

Similarly, we collect subsets based on the same groups from the AGORA dataset. Table 5.5 shows the error on these subsets at different occlusion levels. We observe a similar pattern to the previous experiment. While the *Upper Body*'s occlusion is the most effective, the *Torso* becomes the second. However, for AGORA *Lower Body* is the least important body region in terms of occlusion. As can be seen in Figure 5.9, there is a strong alignment between the effect of *Left Body* and *Right Body*. The statistics on AGORA are in line with expectations of the right and left body of a person have a

similar importance in terms of visibility. The reason may be the uniform distribution of left and right body visibilities across AGORA dataset.

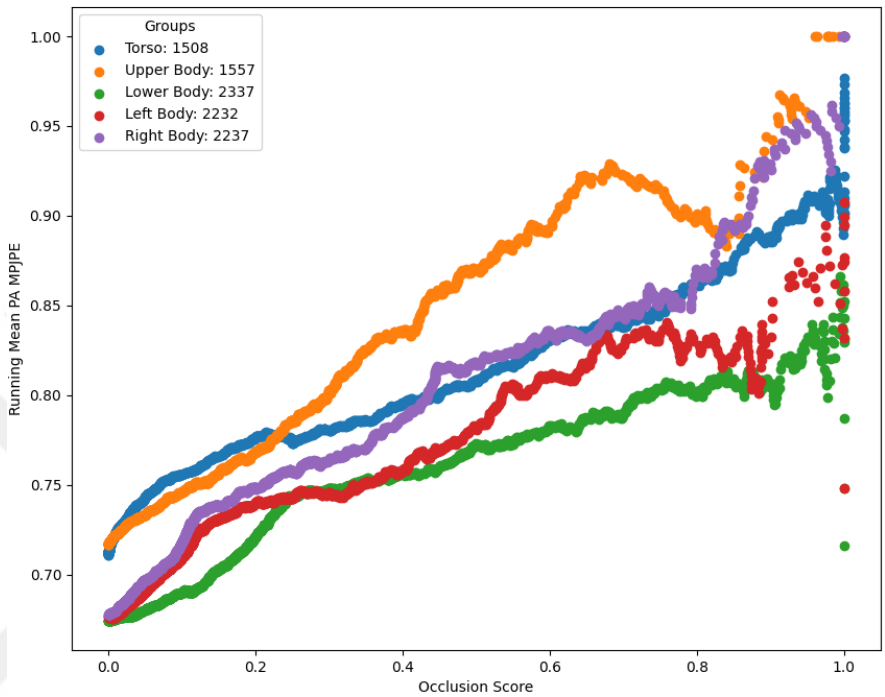

Figure 5.9. ROMP'S PA-MPJPE on handcrafted AGORA groups per occlusion score.

5.3.1. Learning Occlusion Weights per Joint

The groups in Table 5.4 provide binary weights (mask) and have several limitations as described in the section 4.1.2. Custom and more effective weights (coefficients) can be learned from the methods. The framework shown in Figure 4.13 produces a more challenging ranked list for each method depending on the hyperparameter K.

However, to validate the weights, it is necessary to show the transferability of the weights across methods. Therefore, we tested ROMP on the subset formed by the weights learned by BEV. Table 5.6 shows the performance of ROMP on the subset formed by the weights learned from the *most challenging 60 samples* according to BEV (we manually chose K as 60). Figure 5.10 visually demonstrates the Table 5.6. The subset formed by the weights learned from BEV is the most challenging subset among the other hand-crafted subsets, especially at 0.5, 0.6, and 0.7 occlusion scores, which

Table 5.6. Running mean error on 3DPW dataset for learned group.

	Occlusion Score						
Group	**0.2**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Full Body	**0.42**	0.45	0.51	0.65	0.71	0.84	0.71
Torso	0.38	0.42	0.45	0.52	0.58	0.63	0.70
Upper Body	0.39	0.44	0.47	0.59	0.78	**0.87**	**1.0**
Lower Body	0.40	**0.47**	0.52	0.59	0.65	0.56	0.52
Learned Weights	0.41	0.44	**0.53**	**0.73**	**0.82**	0.85	0.76

represents the beginning of heavy occlusion. Besides, the learned group represents the second most challenging group with 0.8 and 0.9 occlusion scores. Figure 5.10 visually verifies this. The weights for each body segment learned from the BEV and validated on ROMP are visualized in Figure 5.11. The joints representing the upper body have a higher weight coefficient than the lower body, aligning the handcrafted version.

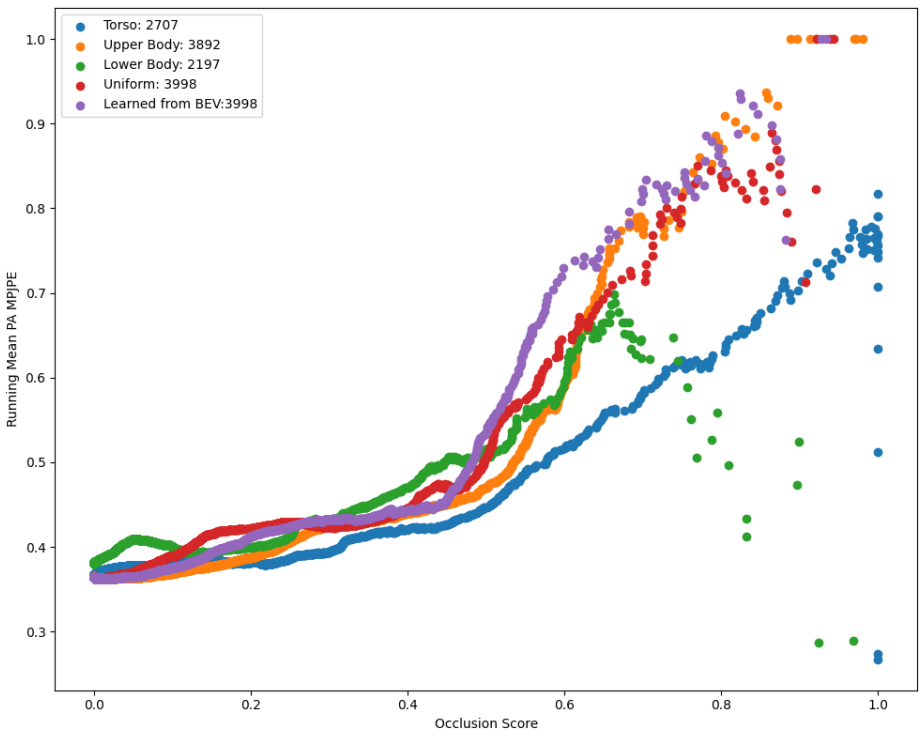

Figure 5.10. ROMP's performance on custom group learned from BEV. Note that the advantage of the learned group is clear for occlusion scores between 0.5 and 0.7.

We also examined the effect of the occlusion of each body segment on the joint error. Figure 5.12 shows the average error on the 3DPW dataset per joint depending on the occlusion of different body segments as a heatmap. The error on the pelvis is important regardless of which body segment is occluded. The joints of the upper body, such as shoulders and elbows, also generally have high errors. This finding is also aligned with the experiments conducted on handcrafted groups, where the highest error is produced on the upper body occluded samples. Note that the joint names on the error (x) axis are listed in Figure 2.1, and the body segments listed on the y-axis are visualized in Figure 4.11, which is used for the Regional Occlusion Index. The method used in this figure is ROMP and the error metric is PA-MPJPE. We observe a similar pattern for AGORA dataset, too. In addition to the pelvis and torso, the hands are also generally affected by the occlusion.

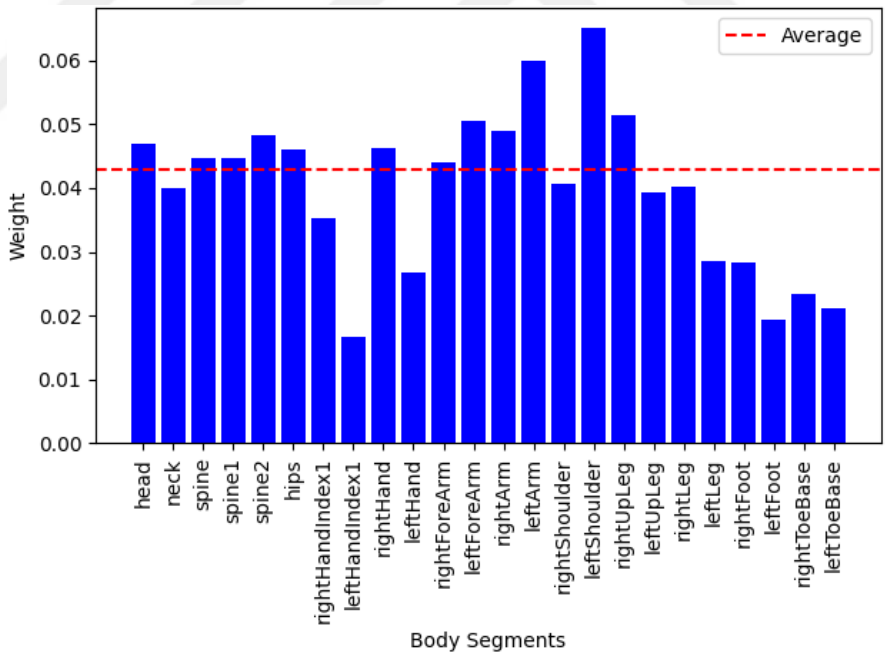

Figure 5.11. Weight of each body segment learned from BEV transferred and validated on ROMP.

5.4. Self Occlusion

We conducted experiments on self-occlusion in two stages. In the first step, we investigated the relationship between the orientation of the body and the error caused

by it. In the second step, we took advantage of the Regional Occlusion Index and quantified the self-occlusion for each body segment.

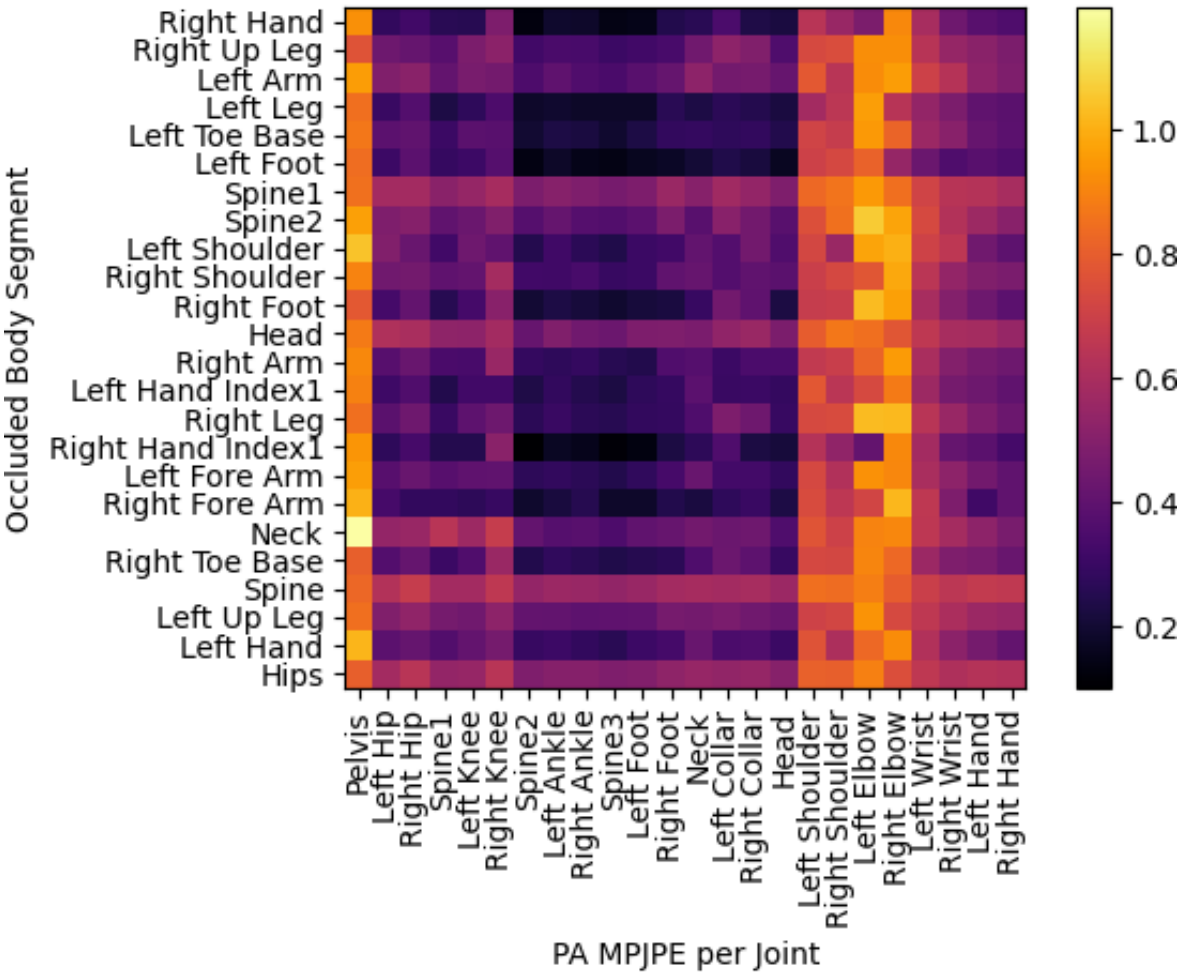

Figure 5.12. Error on each joint according to occlusion per body segment on 3DPW.

5.4.1. Self Occlusion by Body Orientation

For this experiment, all SMPL models in the 3DPW dataset are clustered based on their orientation around the z-axis. Note that, for this experiment, the whole 3DPW dataset is utilized, corresponding to 51K 3D human models approximately. The K-means clustering algorithm is used to divide the dataset into subsets. We manually set the number of clusters to 7 by visually inspecting the results. The degrees under each 3D human body represent the orientation of the person. Note that the clusters are almost equally sampled around the 360-degree unit circle, as expected.

Table 5.7 lists the average error ROMP made per cluster as MPJPE. Note that the errors are symmetrical with respect to the center clusters (Remember cluster centers from Figure 4.14). Model performance tends to degrade as the subject's alignment with the camera decreases. However, the models appear more fragile when the human subject turns away from the camera. Interestingly, the clusters corresponding to 146 and -146 degrees have a rotation offset from the image plane that falls between the clusters corresponding to -99 and 90 degrees. Despite this intermediate positioning, there is a noticeable performance gap between them. This finding reinforces the notion that model vulnerability is influenced not only by the pose of the human subject or the occluded region but also by the specific side of the human facing the camera.

Table 5.7. Body orientation error per cluster.

Orientations	-146°	-99°	-51°	-2°	44°	90°	146°
ROMP	**3.46**	2.99	2.72	2.60	2.72	2.94	**3.42**
BEV	**3.52**	3.05	2.78	2.63	2.73	2.96	**3.46**

5.4.2. Self Occlusion per Body Segment

As described in Section 4.2, single-view 3D HPS estimation inevitably suffers from self-occlusion. Therefore, we tested for self-occlusion in the samples where the methods produced higher error rates. To do this, we used the *outdoors_freestyle_00* scene from 3DPW, which contains only one person in each frame. We did not choose any of the scenes used for multi-person occlusion case because the effect of self-occlusion would be overshadowed due to severe occlusion in those scenes.

Figure 5.13 shows the per-joint PA-MPJPE that ROMP made on the scene. 5.13a shows the error per joint on the whole scene. 5.13b shows the error per joint on the 50 most difficult samples. 5.13c is the normalized version of 5.13b with 5.13a subtracted. Figure 5.13c shows that self-occlusion of the knees, elbows, and wrists most affected performance along with the pelvis. These results are in line with expectations, proving that our Regional Occlusion Index successfully quantifies self-occlusion as well.

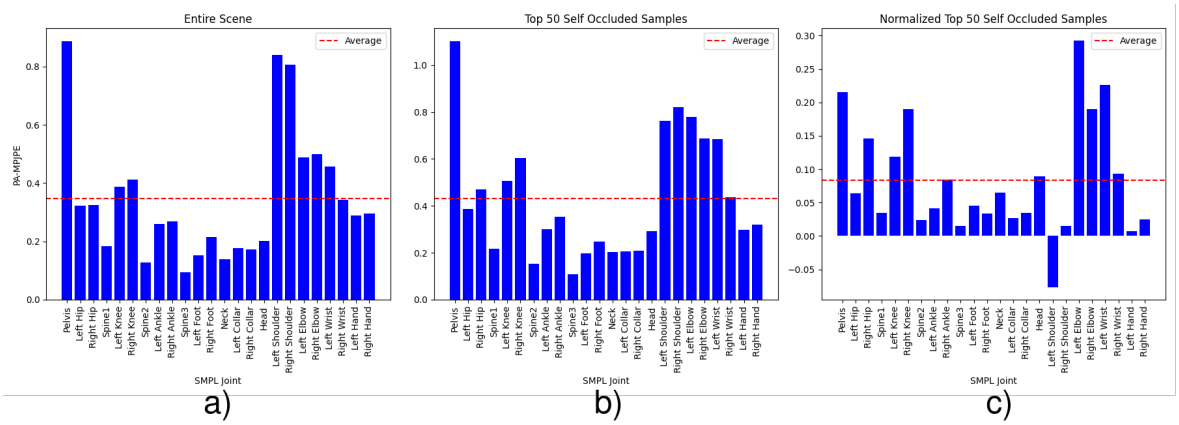

Figure 5.13. Entire scene(a), most challenging top 50(b), and normalized version(c).

5.5. Effect of Modified MPJPE

In Section 4.3, we introduced a modified version of the commonly used metric in 3D HPS MPJPE. Our approach is to evaluate the performance of the model on the visible regions, but not on the occluded regions. (Equation 4.17) Remember that, our proposition is to assess the robustness of models in the case of the presence of occlusion.

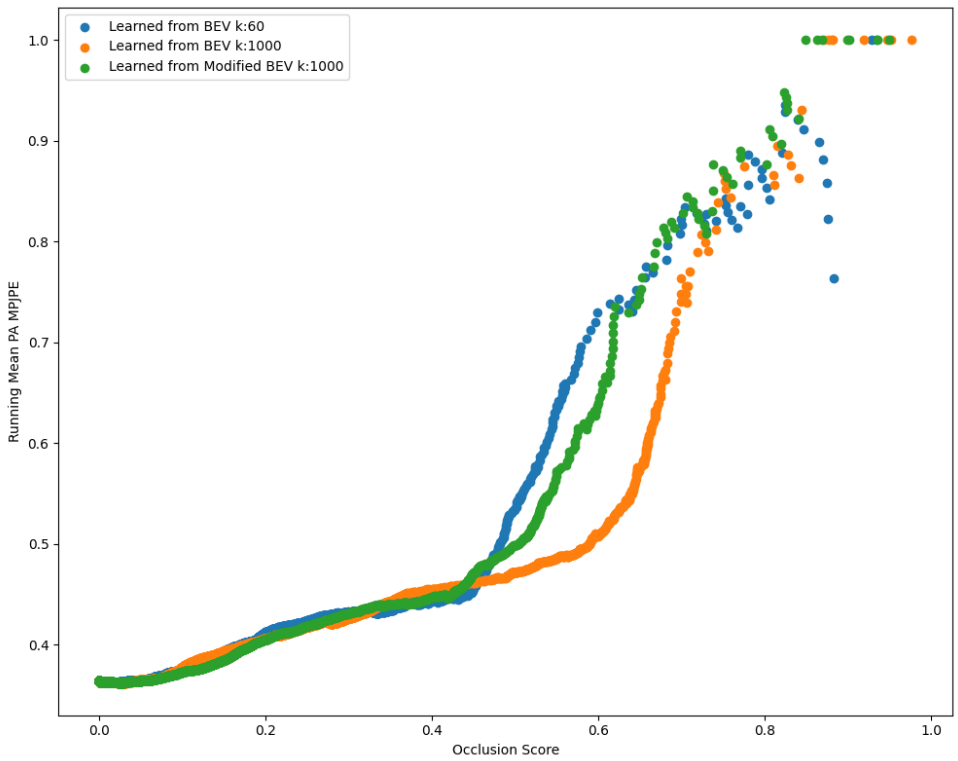

Figure 5.14. Effect of modified MPJPE on learned weights.

We used the *Modified MPJPE* to learn the weights of the models. Recall that in Table 5.6 we reported the best group learned from the 60 most difficult samples for BEV and validated these weights by forming a subset to be tested on ROMP. For this experiment, we replace the error metric used for BEV with the *Modified MPJPE*.

This method failed to improve the baseline when k is 60, but improved for larger values of k. Figure 5.15 compares the baseline where k is 1000 in both the standard and modified versions. While the subset formed by the weights learned from the most difficult top 1000 samples ranked by PA-MPJPE for BEV is quite simple, the one ranked by the modified PA-MPJPE shows a performance closer to the baseline.

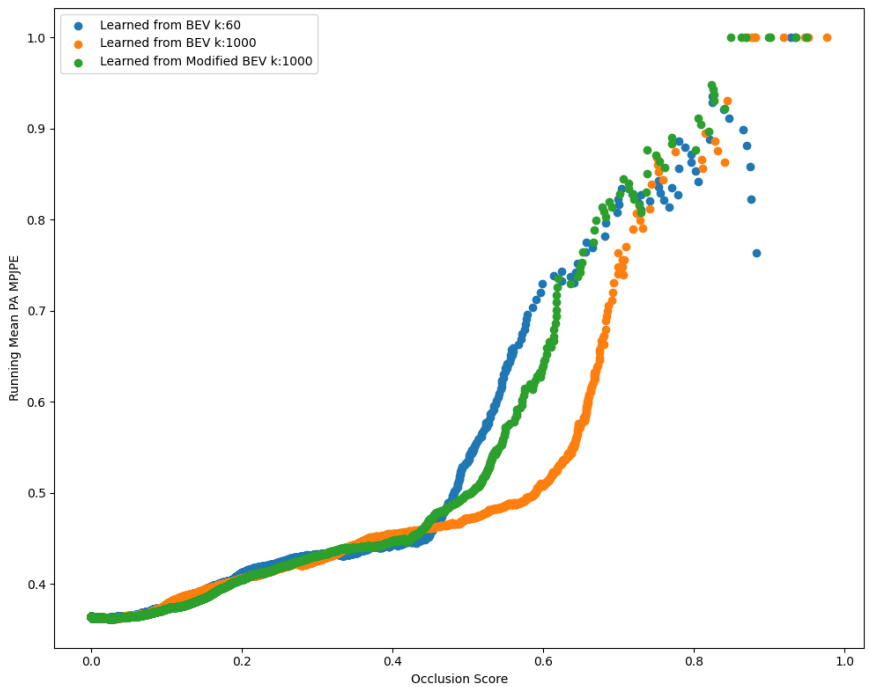

Figure 5.15. Effect of modified MPJPE on learned weights.

Finally, we compared the performance of ROMP and BEV on the subset formed by the Regional Occlusion Index in terms of *Modified PA-MPJPE*. Figure 5.16 shows the difference between the standard and modified versions. In the modified version, a lower error rate is reported for the samples with minimal occlusion. This means that the error occurred when estimating the small and occluded parts, such as the occluded hand or the occluded foot. This finding demonstrates the regularization effect of the modified

MPJPE. It is also consistent with the expectation for the effect of *Modified MPJPE*. If there is a small occlusion, the rest of the body can be reconstructed. However, when there is a large occlusion, the rest of the body is severely affected and the error increases. We see this pattern in the figure. As the occlusion score increases, the error gap between the standard and modified versions closes, and the severe occlusion begins to affect visible regions as well.

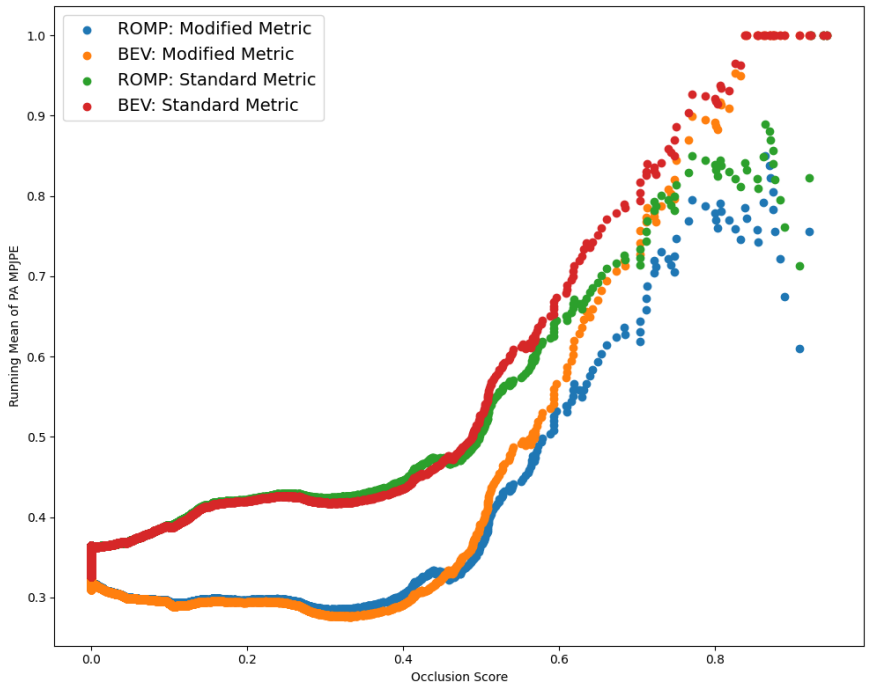

Figure 5.16. Effect of modified MPJPE on model performance assessment on 3DPW dataset.

We also observe a similar pattern in the AGORA dataset. Both models produce less error at the lower occlusion rates under the modified metric, and each method converges to the same point as the occlusion becomes more severe.

6. CONCLUSION

In this study, Occlusion Index is proposed to quantify occlusion in the 3D HPS datasets. The subsets created with the proposed method and the previous method in the literature are compared based on the challenge they introduce to the state-of-the-art occlusion-robust 3D HPS estimation methods.

First, a comprehensive literature review is conducted, including the datasets, methods, and popular methods used in the occlusion-specific studies. Furthermore, the datasets are classified based on the way they were collected and compared to each other, highlighting their strengths and weaknesses in the perspective of occlusion.

Secondly, the drawbacks of the Crowd Index commonly used in the literature are listed with examples. The failure cases are visually demonstrated and the reasons behind these cases are explained in detail.

Third, a novel index for quantifying occlusion, the Occlusion Index, is presented. It is shown with examples that the Occlusion Index overcomes the shortcomings of the Crowd Index. The formulated Occlusion Index has several properties such as *Pseudo Depth Order*, *Pose Invariant Precision*, *Invariance to the Number of Occluders*, *Normalized Score*, and *Detection of Object Occlusions*. These features allow the Occlusion Index to overcome the limitations of the Crowd Index.

Fourth, an even more improved version, the Regional Occlusion Index (ROI), is developed. ROI uses 3D SMPL models as pseudo-instance masks by rendering people in a frame using Z-Buffer algorithm. Therefore, the occlusion of each person is represented by the number of occluded pixels instead of the number of 2D keypoints representing the 2D pose joints including the truncation. This feature allows Regional Occlusion Index to evaluate occlusion in more detail by quantifying it as a real number instead of a binary number.

Fifth, an improved version of the Occlusion Index, called the Body Part Occlusion Index (BPOI), is proposed. The BPOI introduces several types of occlusion depending on the joint being occluded. Since each joint in the body pose carries a different amount of information about the pose itself due to the deformable, non-rigid, and articulated nature of the human pose, the distinction of the joint being occluded is crucial. The human body pose was manually divided into several subgroups based on spatial positions. Regional Occlusion Index also quantifies the occlusion of each body segment separately with a real number by rendering them one at a time. Therefore, BPOI enabled body part-specific occlusion subsets to determine the sensitivity to occlusion of each region.

Sixth, an innovative framework is described that allows the importance of each body segment in terms of occlusion to be learned from the methods instead of manually filtering.

Seventh, ROI is extended to the case of self-occlusion, which is an unavoidable problem in single-view 3D HPS. The occlusion of each body segment by another segment of the same person is quantified, and the effect of each segment is also quantified.

Eighth, a modified version of the commonly used evaluation metric MPJPE is mathematically proposed. The modified version evaluates only the visible regions instead of calculating over all joints regardless of occlusion. This new modified metric introduces a regularization effect and puts more emphasis on the 3D reconstruction of the visible regions despite the presence of occlusion in the sample. Therefore, the small errors caused by the small occluded regions are pruned and the robustness to the severe occlusions is more promoted.

Finally, the superiority of the proposed novel index over the existing one and the use cases revealing the intrinsic properties of the 3D HPS estimators are demonstrated with the experiments. The experiments quantitatively and qualitatively support the propositions of this study.

7. FUTURE WORKS

This study proposes an improved version of the existing occlusion quantification metric Crowd Index. Throughout the thesis, the proposed index is strengthened with different aspects and the ability to successfully quantify occlusions is demonstrated. However, there is still room for improvement.

Context is one of the key factors leading to human pose expectations. A context-aware approach would achieve better quantification by assessing whether the occluded parts are trivial given the context. Since occlusion quantification is more critical for single-view reconstruction, a *Action Recognition in Still Images* method can classify the person's action and dynamically adapt weights to the visibility of joints.

In this study, the amount of occlusion is quantified by the Regional Occlusion Index and the weight is learned from the output of a pre-trained 3D HPS estimator over the dataset. However, a more sophisticated approach would be to integrate the progress of the weight regression into the training process of an occlusion robust 3D HPS estimator and report the performance of the model as a function of the current weight learned at that iteration. This approach not only provides more accurate weights for occlusion quantification but also improves the occlusion robustness of the method.

Although our pseudo instance masks rendered from SMPL models work pretty well, they do not remove the necessity for occlusion-robust instance segmentation models. Such a technique would improve the Occlusion Index more.

REFERENCES

1. Joo, H., H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara and Y. Sheikh, "Panoptic Studio: A Massively Multiview System for Social Motion Capture", ArXiv 1612.03153 [cs], 2015.

2. Choi, H., G. Moon and K. M. Lee, "Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose", *The Proceedings of the European Conference on Computer Vision*, Glasgow, Scotland, pp. 769–787, 2020.

3. Choi, H., G. Moon, J. Park and K. M. Lee, "Learning to Estimate Robust 3D Human Mesh from In-the-wild Crowded Scenes", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, New Orleans, Louisiana, USA, pp. 1475–1484, 2022.

4. Kanazawa, A., M. J. Black, D. W. Jacobs and J. Malik, "End-to-End Recovery of Human Shape and Pose", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, pp. 7122–7131, 2018.

5. Zhang, T., B. Huang and Y. Wang, "Object-Occluded Human Shape and Pose Estimation from a Single Color Image", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Virtual, pp. 7376–7385, 2020.

6. Bogo, F., A. Kanazawa, C. Lassner, P. Gehler, J. Romero and M. J. Black, "Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image", *The Proceedings of the European Conference on Computer Vision*, Amsterdam, Netherlands, pp. 561–578, 2016.

7. Kocabas, M., N. Athanasiou and M. J. Black, "VIBE: Video Inference for Human Body Pose and Shape Estimation", *The Proceedings of the Conference on*

Computer Vision and Pattern Recognition, Virtual, pp. 5253–5263, 2020.

8. Yu, Z., J. S. Yoon, I. K. Lee, P. Venkatesh, J. Park, J. Yu and H. S. Park, "HUMBI: A Large Multiview Dataset of Human Body Expressions", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Virtual, pp. 2990–3000, 2020.

9. Jiang, H., J. Cai and J. Zheng, "Skeleton-Aware 3D Human Shape Reconstruction from Point Clouds", *The Proceedings of the International Conference on Computer Vision*, Seoul, South Korea, pp. 5431–5441, 2019.

10. Yu, T., Z. Zheng, K. Guo, P. Liu, Q. Dai and Y. Liu, "Function4D: Real-Time Human Volumetric Capture from Very Sparse Consumer RGB-D Sensors", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Virtual, pp. 5746–5756, 2021.

11. Zheng, Y., R. Shao, Y. Zhang, T. Yu, Z. Zheng, Q. Dai and Y. Liu, "Deepmulticap: Performance Capture of Multiple Characters Using Sparse Multiview Cameras", *The Proceedings of the International Conference on Computer Vision*, Virtual, pp. 6239–6249, 2021.

12. Yuan, Y., U. Iqbal, P. Molchanov, K. Kitani and J. Kautz, "GLAMR: Global Occlusion-Aware Human Mesh Recovery with Dynamic Cameras", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, New Orleans, Louisiana, USA, pp. 11038–11049, 2022.

13. Liang, J. and M. C. Lin, "Shape-Aware Human Pose and Shape Reconstruction Using Multi-View Images", *The Proceedings of the International Conference on Computer Vision*, Seoul, South Korea, pp. 4352–4362, 2019.

14. Zhang, H., Y. Tian, X. Zhou, W. Ouyang, Y. Liu, L. Wang and Z. Sun, "PyMAF: 3D Human Pose and Shape Regression with Pyramidal Mesh Alignment

Feedback Loop", *The Proceedings of the International Conference on Computer Vision*, Virtual, pp. 11446–11456, 2021.

15. Sun, Y., W. Liu, Q. Bao, Y. Fu, T. Mei and M. J. Black, "Putting People in Their Place: Monocular Regression of 3D People in Depth", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, New Orleans, Louisiana, USA, pp. 13243–13252, 2022.

16. Mehta, D., O. Sotnychenko, F. Mueller, W. Xu, S. Sr idhar, G. Pons-Moll and C. Theobalt, "Single-Shot Multi-Person 3D Pose Estimation from Monocular RGB", *International Conference on 3D Vision*, Verona, Italy, pp. 120–130, 2018.

17. Sun, Y., Q. Bao, W. Liu, Y. Fu, M. J. Black and T. Mei, "Monocular, One-stage, Regression of Multiple 3D People", *The Proceedings of the International Conference on Computer Vision*, Virtual, pp. 11179–11188, 2021.

18. Kolotouros, N., G. Pavlakos, M. J. Black and K. Daniilidis, "Learning to Reconstruct 3D Human Pose and Shape via Model Fitting in the Loop", *The Proceedings of the International Conference on Computer Vision*, Seoul, South Korea, pp. 2252–2261, 2019.

19. Zheng, Z., T. Yu, Y. Wei, Q. Dai and Y. Liu, "DeepHuman: 3D Human Reconstruction From a Single Image", *The Proceedings of the International Conference on Computer Vision*, Seoul, South Korea, pp. 7739–7749, 2019.

20. Mihajlovic, M., S. Saito, A. Bansal, M. Zollhoefer and S. Tang, "COAP: Compositional Articulated Occupancy of People", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, New Orleans, Louisiana, USA, pp. 13201–13210, 2022.

21. Tian, Y., H. Zhang, Y. Liu and L. Wang, "Recovering 3D Human Mesh from Monocular Images: A Survey", *Transactions on Pattern Analysis and Machine*

Intelligence, Vol. 45, No. 12, p. 15406–15425, 2023.

22. Khirodkar, R., S. Tripathi and K. Kitani, "Occluded Human Mesh Recovery", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, New Orleans, Louisiana, USA, pp. 1715–1725, 2022.

23. Zhang, J., D. Yu, J. H. Liew, X. Nie and J. Feng, "Body Meshes as Points", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Virtual, pp. 546–556, 2021.

24. Zhou, Q., S. Wang, Y. Wang, Z. Huang and X. Wang, "Human De-occlusion: Invisible Perception and Recovery for Humans", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Virtual, pp. 3691–3701, 2021.

25. Joo, H., N. Neverova and A. Vedaldi, "Exemplar Fine-Tuning for 3D Human Model Fitting Towards In-the-wild 3D Human Pose Estimation", *International Conference on 3D Vision*, Prague, Czechia, pp. 42–52, 2021.

26. Yu, Z., J. Wang, J. Xu, B. Ni, C. Zhao, M. Wang and W. Zhang, "Skeleton2Mesh: Kinematics Prior Injected Unsupervised Human Mesh Recovery", *The Proceedings of the International Conference on Computer Vision*, Virtual, pp. 8619–8629, 2021.

27. Rafi, U., J. Gall and B. Leibe, "A Semantic Occlusion Model for Human Pose Estimation from a Single Depth Image", *The Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops*, Boston, Massachusetts, USA, pp. 67–74, 2015.

28. Kocabas, M., C.-H. P. Huang, O. Hilliges and M. J. Black, "PARE: Part Attention Regressor for 3D Human Body Estimation", *The Proceedings of the International Conference on Computer Vision*, Virtual, pp. 11127–11137, 2021.

29. Patel, P., C.-H. P. Huang, J. Tesch, D. T. Hoffmann, S. Tripathi and M. J. Black, "AGORA: Avatars in Geography Optimized for Regression Analysis", *The Pro-*

ceedings of the Conference on Computer Vision and Pattern Recognition, Virtual, pp. 13468–13478, 2021.

30. Von Marcard, T., R. Henschel, M. J. Black, B. Rosenhahn and G. Pons-Moll, "Recovering Accurate 3D Human Pose In-the-wild Using IMUs and a Moving Camera", *The Proceedings of the European Conference on Computer Vision*, Munich, Germany, pp. 601–617, 2018.

31. Zhang, S.-H., R. Li, X. Dong, P. Rosin, Z. Cai, X. Han, D. Yang, H. Huang and S.-M. Hu, "Pose2Seg: Detection Free Human Instance Segmentation", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Long Beach, California, USA, pp. 889–898, 2019.

32. Black, M. J., P. Patel, J. Tesch and J. Yang, "BEDLAM: A Synthetic Dataset of Bodies Exhibiting Detailed Lifelike Animated Motion", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 8726–8737, 2023.

33. Ionescu, C., D. Papava, V. Olaru and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments", *Transactions on Pattern Analysis and Machine Intelligence*, Vol. 36, No. 7, pp. 1325–1339, 2013.

34. Trumble, M., A. Gilbert, C. Malleson, A. Hilton and J. Collomosse, "Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors", *British Machine Vision Conference*, London, UK, 2017.

35. Cai, Z., M. Zhang, J. Ren, C. Wei, D. Ren, J. Li, Z. Lin, H. Zhao, S. Yi, L. Yang, C. C. Loy and Z. Liu, "Playing for 3D Human Recovery", ArXiv 2110.07588 [cs], 2021.

36. Sigal, L., A. O. Balan and M. J. Black, "Humaneva: Synchronized Video and

Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion", *International Journal of Computer Vision*, Vol. 87, No. 1-2, pp. 4–27, 2010.

37. Mehta, D., H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu and C. Theobalt, "Monocular 3D Human Pose Estimation In-the-wild Using Improved CNN Supervision", *International Conference on 3D Vision*, Qingdao, China, pp. 506–516, 2017.

38. Johnson, S. and M. Everingham, "Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation", *British Machine Vision Conference*, Aberystwyth, UK, p. 5, 2010.

39. Girgin, E., B. Gökberk and L. Akarun, "A Novel Occlusion Index", *Signal Processing and Communications Applications Conference*, İstanbul, Türkiye, pp. 1–4, 2023.

40. Loper, M., N. Mahmood, J. Romero, G. Pons-Moll and M. J. Black, "SMPL: A Skinned Multi-Person Linear Model", *Transactions on Graphics*, Vol. 34, No. 6, pp. 1–16, 2015.

41. Pearson, K., "LIII. On Lines and Planes of Closest Fit to Systems of Points in Space", *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Vol. 2, No. 11, pp. 559–572, 1901.

42. Loper, M., N. Mahmood and M. J. Black, "MoSh: Motion and Shape Capture from Sparse Markers", *Transactions on Graphics*, Vol. 33, No. 6, pp. 1–13, 2014.

43. Pavlakos, G., V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas and M. J. Black, "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Long Beach, California, USA, pp. 10975–10985, 2019.

44. Mahmood, N., N. Ghorbani, N. F. Troje, G. Pons-Moll and M. J. Black, "AMASS: Archive of Motion Capture as Surface Shapes", *The Proceedings of the International Conference on Computer Vision*, Seoul, South Korea, pp. 5442–5451, 2019.

45. Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick, "Microsoft COCO: Common Objects in Context", *The Proceedings of the European Conference on Computer Vision*, Zurich, Switzerland, pp. 740–755, 2014.

46. Fang, Q., Q. Shuai, J. Dong, H. Bao and X. Zhou, "Reconstructing 3D Human Pose by Watching Humans in the Mirror", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Virtual, pp. 12814–12823, 2021.

47. Sengupta, A., I. Budvytis and R. Cipolla, "Synthetic Training for Accurate 3D Human Pose and Shape Estimation In-the-wild", ArXiv 2009.10013 [cs], 2020.

48. Johnson, S. and M. Everingham, "Learning Effective Human Pose Estimation from Inaccurate Annotation", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Colorado Springs, Colorado, USA, pp. 1465–1472, 2011.

49. Andriluka, M., U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall and B. Schiele, "PoseTrack: A Benchmark for Human Pose Estimation and Tracking", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, pp. 5167–5176, 2018.

50. Lassner, C., J. Romero, M. Kiefel, F. Bogo, M. J. Black and P. V. Gehler, "Unite the People: Closing the Loop Between 3D and 2D Human Representations", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, pp. 6050–6059, 2017.

51. Andriluka, M., L. Pishchulin, P. Gehler and B. Schiele, "2D Human Pose Estimation: New Benchmark and State of the Art Analysis", *The Proceedings of the*

Conference on Computer Vision and Pattern Recognition, Columbus, Ohio, USA, pp. 3686–3693, 2014.

52. Li, Z., T. Dekel, F. Cole, R. Tucker, N. Snavely, C. Liu and W. T. Freeman, "Learning the Depths of Moving People by Watching Frozen People", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Long Beach, California, USA, pp. 4521–4530, 2019.

53. Peng, S., Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao and X. Zhou, "Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Virtual, pp. 9054–9063, 2021.

54. Muller, L., A. A. Osman, S. Tang, C.-H. P. Huang and M. J. Black, "On Self-Contact and Human Pose", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Virtual, pp. 9990–9999, 2021.

55. Varol, G., J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev and C. Schmid, "Learning from Synthetic Humans", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, pp. 109–117, 2017.

56. Zanfir, A., E. Marinoiu and C. Sminchisescu, "Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes: The Importance of Multiple Scene Constraints", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, pp. 2148–2157, 2018.

57. Popa, A.-I., M. Zanfir and C. Sminchisescu, "Deep Multitask Architecture for Integrated 2D and 3D Human Sensing", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, pp. 6289–6298, 2017.

58. Kolotouros, N., G. Pavlakos and K. Daniilidis, "Convolutional Mesh Regression for Single-Image Human Shape Reconstruction", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Long Beach, California, USA, pp. 4501–4510, 2019.

59. Kipf, T. N. and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks", ArXiv 1609.02907 [cs], 2016.

60. Cho, K., B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation", ArXiv 1406.1078 [cs], 2014.

61. Huang, J.-B. and M.-H. Yang, "Estimating Human Pose from Occluded Images", *Asian Conference on Computer Vision*, Xi'an, China, pp. 48–60, 2009.

62. Sárándi, I., T. Linder, K. O. Arras and B. Leibe, "How Robust is 3D Human Pose Estimation to Occlusion?", ArXiv 1808.09316 [cs], 2018.

63. Everingham, M., L. Van Gool, C. K. Williams, J. Winn and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge", *International Journal of Computer Vision*, Vol. 88, pp. 303–338, 2010.

64. Microsoft Corporation, "Kinect for Windows SDK 2.0", `https://developer.microsoft.com/en-us/windows/kinect`, accessed on July 10, 2023.

65. Jiang, W., N. Kolotouros, G. Pavlakos, X. Zhou and K. Daniilidis, "Coherent Reconstruction of Multiple Humans from a Single Image", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Virtual, pp. 5579–5588, 2020.

66. Girshick, R., J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation", *The Proceedings of the*

Conference on Computer Vision and Pattern Recognition, Columbus, Ohio, USA, pp. 580–587, 2014.

67. Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is All you Need", ArXiv 1706.03762 [cs], 2017.

68. Ioffe, S. and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", *International Conference on Machine Learning*, Lille, France, pp. 448–456, 2015.

69. Li, J., C. Wang, H. Zhu, Y. Mao, H.-S. Fang and C. Lu, "Crowdpose: Efficient Crowded Scenes Pose Estimation and a New Benchmark", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Long Beach, California, USA, pp. 10863–10872, 2019.

70. Hesse, N., S. Pujades, J. Romero, M. J. Black, C. Bodensteiner, M. Arens, U. G. Hofmann, U. Tacke, M. Hadders-Algra, R. Weinberger *et al.*, "Learning an Infant Body Model from RGB-D Data for Accurate Full Body Motion Analysis", *International Conference on Medical Image Computing and Computer Assisted Intervention*, Granada, Spain, pp. 792–800, 2018.

71. Cho, H., Y. Cho, J. Ahn and J. Kim, "Implicit 3D Human Mesh Recovery using Consistency with Pose and Shape from Unseen-view", *The Proceedings of the Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 21148–21158, 2023.

72. Straßer, W., *Zukünftige Arbeiten. Schnelle Kurven-und Flächendarstellung auf grafischen Sichtgeräten*, Ph.D. Thesis, Technical University of Berlin, 1974.

73. Lloyd, S., "Least Squares Quantization in PCM", *Transactions on Information Theory*, Vol. 28, No. 2, pp. 129–137, 1982.

74. Gower, J. C. and G. B. Dijksterhuis, "Procrustes Problems", *Oxford Statistical*

Science Series, Vol. 30, No. 4, pp. 799–801, 2004.

APPENDIX A: ERRONEOUS ANNOTATIONS

A.1. Erroneous Annotations in 3DPW

In Figure A.1, some of the erroneous annotations from 3DPW are demonstrated. A.1a and c show the images, and b and d show the rendered masks of 3D models. Observe the bizarre on the mask indicated with red.

List of the erroneously annotated frames we found:

- courtyard_dancing_01/image_00114.jpg
- courtyard_dancing_01/image_00106.jpg
- courtyard_dancing_01/image_00107.jpg
- courtyard_dancing_01/image_00118.jpg
- courtyard_dancing_01/image_00108.jpg
- courtyard_dancing_01/image_00116.jpg
- courtyard_dancing_01/image_00113.jpg
- courtyard_dancing_01/image_00119.jpg
- courtyard_dancing_01/image_00117.jpg
- courtyard_dancing_01/image_00120.jpg
- courtyard_dancing_01/image_00111.jpg
- courtyard_dancing_01/image_00105.jpg
- courtyard_dancing_01/image_00115.jpg
- courtyard_dancing_01/image_00109.jpg
- courtyard_dancing_01/image_00112.jpg
- courtyard_dancing_01/image_00110.jpg
- courtyard_dancing_01/image_00121.jpg
- courtyard_dancing_01/image_00104.jpg
- courtyard_dancing_01/image_00103.jpg
- courtyard_dancing_01/image_00122.jpg

Figure A.1. Some of the erroneous annotations (b, d) and corresponding images (a, c) in 3DPW

- courtyard_hug_00/image_00177.jpg
- courtyard_hug_00/image_00216.jpg
- courtyard_hug_00/image_00217.jpg

A.2. Erroneous Annotations in OCHuman

In Figure A.2, some of the erroneous annotations from OCHuman are demonstrated.

List of the erroneously annotated frames we found:

- 004163.jpg
- 003285.jpg
- 001703.jpg
- 000394.jpg
- 004240.jpg
- 003285.jpg
- 002817.jpg
- 003955.jpg
- 003548.jpg
- 000624.jpg
- 004048.jpg
- 000856.jpg

Figure A.2. Some of the erroneous annotations in OCHuman