VOccl3D: A Video Benchmark Dataset for 3D Human Pose and Shape Estimation under real Occlusions

Yash Garg, Saketh Bachu, Arindam Dutta, Rohit Lal†, Sarosij Bose,
Calvin-Khang Ta‡, M. Salman Asif, Amit Roy-Chowdhury
University of California, Riverside
{ygarg002,sbach008,adutt020,rlal011,sbose007,cta003,sasif,amitrc}@ucr.edu

Figure 1. We propose **VOccl3D**, a large-scale synthetic video dataset specifically designed for training and evaluating algorithms for 3D human pose and shape estimation (HPS) in realistic occlusion scenarios. VOccl3D comprises over 250,000 frames, with a total runtime exceeding 2 hours and 30 minutes. Compared to previous occlusion-based datasets, VOccl3D has diverse body shapes, textures, and most importantly, significant and realistic occlusions within the scenes. In addition to body shape and pose, VOccl3D provides other necessary ground truth information, such as bounding boxes, body part segmentations, and human silhouettes. The top row illustrates various frames from our dataset, showcasing diverse occlusions, clothing textures, and motions, while the bottom row represents a sequence of frames from a video sequence within our dataset.

Abstract

*Human pose and shape (HPS) estimation methods have been extensively studied, with many demonstrating high zero-shot performance on in-the-wild images and videos. However, these methods often struggle in challenging scenarios involving complex human poses or significant occlusions. Although some studies address 3D human pose estimation under occlusion, they typically evaluate performance on datasets that lack realistic or substantial occlusions, e.g., most existing datasets introduce occlusions with random patches over the human or clipart-style overlays, which may not reflect real-world challenges. To bridge this gap in realistic occlusion datasets, we introduce a novel benchmark dataset, VOccl3D, a **V**ideo-based human **Occ**lusion dataset with **3D** body pose and shape annotations. Inspired by works such as AGORA and BEDLAM, we constructed this dataset using advanced computer graphics rendering techniques, incorporating diverse real-world occlusion scenarios, clothing textures, and human motions. Additionally, we fine-tuned recent HPS methods, CLIFF and BEDLAM-CLIFF, on our dataset, demonstrating significant qualitative and quantitative improvements across multiple public datasets, as well as on the test split of our dataset, while comparing its performance with other state-of-the-art methods. Furthermore, we leveraged our dataset to enhance human detection performance under occlusion by fine-tuning an existing object detector, YOLO11, thus leading to a robust end-to-end HPS estimation system under occlusions. Overall, this dataset serves as a valuable resource for future research aimed at benchmarking methods designed to handle occlusions, offering a more realistic alternative to existing occlusion datasets. See the Project page for code and dataset:https://yashgarg98.github.io/VOccl3D-dataset/*

† Currently at NASA MSFC IMPACT. ‡ Currently at Dolby Laboratories. Work done while the authors were at UCR.

1. Introduction

Monocular 3D human pose and shape (HPS) estimation is a complex yet essential task in computer vision. It has applications in surveillance [8, 20], autonomous robotics [3, 46, 56], human motion analysis [10, 11], and clinical assessment [9, 26, 37]. Since the introduction of the neural network-based Human Mesh Recovery Network (HMR) [21], numerous approaches have advanced the field by enabling accurate estimation of SMPL[32] pose and shape from a single RGB image. Recent methods [2, 24, 27, 28, 33, 48] demonstrate impressive zero-shot HPS performance on in-the-wild RGB images and video sequences. However, these methods still face limitations in achieving high performance in challenging scenarios, such as complex human poses and under significant occlusions.

Achieving robust HPS under occlusion remains a challenging problem due to the contextual ambiguity of the occluded body parts. Several methods [25, 44, 55, 59] have been proposed to address the challenges caused by occlusions. Methods like PARE [25] focus on using visible body parts to infer occluded areas, enhancing estimation accuracy with partial views. Temporal models like HuMoR [44] and GLAMR [55] use generative frameworks to ensure pose continuity over time. HuMoR predicts pose distributions, while GLAMR incorporates global trajectory data to fill in missing poses. Recent methods, like STRIDE [27], leverage a large-scale pre-trained motion prior model to achieve temporally coherent pose reconstruction under occlusions. While these methods effectively leverage visible information and temporal continuity, they still struggle under severe occlusions due to limited exposure to occluded scenarios in their training data.

Evaluations of the existing methods are often limited to occluded datasets that lack scene diversity [38], involve moderate or lower levels of occlusion [16, 49], or use patch-based occlusion datasets for training and inference [59], as illustrated in Figure 2. Most importantly, there remains an urgent need for a realistic, diverse, and significantly occluded human pose and shape dataset to advance the task of 3D HPS estimation. To address this gap, our work proposes VOccl3D, a large-scale video dataset with synthetic humans in real scenes for 3D human pose and shape estimation. Following prior works such as BEDLAM [4] and Synthmocap [14], we use an advanced computer graphics engine to render a highly realistic synthetic dataset. Prior research has shown that fine-tuning models on synthetic datasets can significantly improve HPS estimation performance on real-world datasets [4]. Synthetic datasets provide "perfect" ground-truth annotations and eliminate the need for costly sensors, making it computationally efficient to generate large-scale datasets.

Our proposed dataset includes approximately 250,000 images and 400 video sequences. To introduce diversity, we incorporate human motion samples from the AMASS [34] dataset, over 200 distinct clothing textures for sequences containing both male and female genders, and 40 real background scenes with occlusions. Additionally, we provide occlusion labels for each body joint. Unlike prior methods [4, 39] which render synthetic environments, we utilize 3D Gaussian splatting (3DGS) [22] to create 3D representations of background scenes. This approach enhances the realism of rendered scenes, making them closely resemble real-world captured data. Reconstructing background scenes using 3DGS offers the flexibility to create domain-specific datasets, such as for agriculture, street or indoor environments using raw RGB video captures without relying on costly or limited graphic assets. Figure 1 showcases samples from our dataset, illustrating the diversity in clothing, environments, human motions, and occlusions.

We fine-tuned the CLIFF [28] and BEDLAM-CLIFF [4] models on our dataset to demonstrate that training with synthetic data enhances the performance of existing HPS estimation methods under occlusion. We evaluated state-of-the-art HPS estimation methods on the test split of our dataset across three occlusion levels: high, medium, and low. The results show that existing methods perform poorly under high occlusion, whereas our fine-tuned models achieve significant improvements. To assess performance on real-world data, we perform evaluations on the 3DPW and OCMotion datasets, which contain low-level occlusions. Additionally, we created a variant of the 3DPW dataset by adding black patches to evaluate the robustness

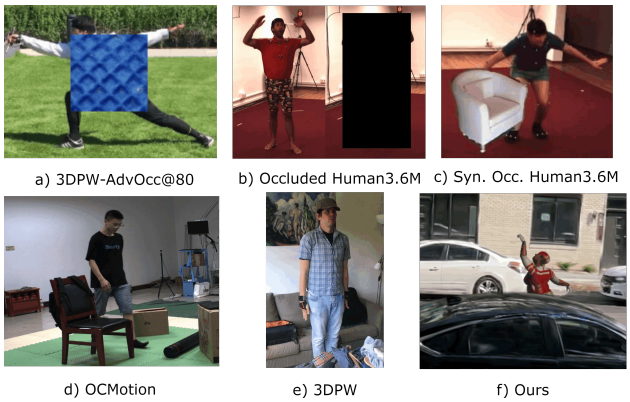

Figure 2. Samples from various occlusion-based HPS datasets. The **top row** displays datasets with artificial, patch-based occlusions applied to existing datasets, including 3DPW-AdvOcc [49, 59], Occluded Human3.6M [18, 27], and Synthetic Occlusion Human3.6M [18, 58]. The **bottom row** presents samples from two datasets with natural occlusions: OCMotion [16] and 3DPW [49]. Notably, the top-row images exhibit unrealistic occlusions, which lack the realism of naturally occurring occlusions. The bottom row images contain natural occlusions but are sparse and infrequent. Compared to existing datasets, our proposed VOccl3D dataset features more realistic occlusions.

against high-occlusion real scenarios. Our fine-tuned models outperformed recent HPS methods on these real-world datasets. Furthermore, our experiments highlight the importance of an effective human object detector for improving HPS performance under occlusion. To address this, we fine-tuned the YOLO11[19] object detector using our dataset, enhancing its performance under occluded scenarios. VOccl3D is available for researchers to benchmark and evaluate occlusion-aware methods.

In summary, we make the following key contributions:

1. We propose **VOccl3D**, a novel large-scale, realistic video-based dataset of occluded synthetic humans in real scenes for 3D human pose and shape estimation.
2. We demonstrate both qualitatively and quantitatively that existing HPS estimation methods fine-tuned with VOccl3D outperform other methods on real-world datasets with occlusions.
3. We improved the performance of the human object detector under occlusion scenarios by leveraging the VOccl3D dataset which is a crucial component for a robust HPS estimation.
4. VOccl3D can serve as a benchmark dataset for evaluating methods specifically designed to perform under occlusion across various tasks, including human and body-part segmentation, 2D/3D pose estimation, and human bounding box detection.

2. Related Works

Synthetic Data Generation for Human Pose Estimation. Human pose estimation is crucial in computer vision, but state-of-the-art methods depend on costly, labor-intensive labeled datasets. Several notable synthetic datasets have advanced research in human pose estimation. SynBody [53] provides a large-scale synthetic dataset for human mesh recovery and view synthesis, while RePoGen [42] enables fine-grained control over pose and viewpoint to generate rare, complex poses. Human3.6m [18] provides additional small mixed reality dataset by inserting 3D animation models with background scenes. BEDLAM [4] further highlights that models trained solely on synthetic data can outperform real-data-trained counterparts, emphasizing the importance of high-quality synthetic datasets for transferable models. Further, PressurePose [7] simulates interactions between articulated and soft bodies, capturing fine-grained contact dynamics. SynthMocap [14] extends this by providing expressive synthetic data with detailed body, hand, and face movements. Despite their success, these synthetic datasets lack emphasis on significant occlusions, a key real-world challenge. Addressing this limitation, our work proposes a synthetic dataset tailored to 3D pose estimation under heavy and realistic occlusion, aiming to bridge this gap and enhance robustness in occlusion-prone environments.

Image and Video based Human Pose Estimation. Estimating 2D/3D pose and shape from single RGB image has widely been explored in previous research [4, 28, 47, 51, 54]. Methods such as [47] use a conditional variational autoencoder for 2D-to-3D lifting in pose estimation. The pose estimation approach proposed in [51] applies normalizing flows [45] for 2D-3D mapping. Large-scale datasets like BEDLAM [4] have improved pose estimation models like CLIFF [28] and HMRNet [21]. However, these models struggle with generalizing to unseen scenarios with severe occlusions due to limited training on such cases.

In addition, video-based human pose estimation has significantly improved challenging datasets. Early works like [60] use the EM algorithm to estimate the 3D pose from monocular video through 2D joint uncertainty maps. The method in [41] employs dilated temporal convolutions and semi-supervised learning for 3D pose estimation, while [2] uses the SMPL model [32] to extract pose and shape parameters, refining models like HMR with bundle adjustment. VIBE [24] applies adversarial learning with AMASS [34] for 3D pose extraction, while MEVA [33] improves accuracy and smoothness using a variational autoencoder to address VIBE's high acceleration error. Owing to these drawbacks, HuMoR employs a conditional variational autoencoder for robust pose estimation, while the state-of-the-art method WHAM [48] integrates 2D-3D lifting and SLAM for accurate global motion estimation. While these methods perform well on various datasets, studies like [27] highlight performance drops under significant occlusions, as most datasets contain only sparse occlusions, making models struggle with unseen heavy or prolonged occlusions.

Human Pose Estimation under Occlusion. Despite significant progress in HPS estimation, handling occlusions remains a major challenge. This is because, in 3D pose estimation, missing depth cues make reconstruction far more ambiguous than in 2D. Early works like [6] used data augmentation with occlusion labels to enhance robustness in pose estimation. Latest methods like GLAMR [55] use a deep generative motion infiller to handle missing poses under severe occlusions and a global optimization framework to refine motion trajectories. Additionally, some methods [23, 33, 57] approach the issue of missing poses as a pose refinement problem, leveraging temporal smoothness to address it effectively. SmoothNet [57] introduced a temporal motion refinement network for refining the poses obtained from the image-based pose estimators to alleviate jitters. These methods perform well under sparse, infrequent occlusions across frames but struggle with natural, prolonged occlusions, as they lack training for such instances. All these works evaluate their methods on datasets with sparse occlusions, as no existing datasets contain significant occlusions. To address this gap, we introduce VOccl3D, a novel dataset designed specifically for 3D HPS estimation under significant occlusions.

3. VOccl3D Dataset

In this section, we outline the key components required to construct the VOccl3D dataset. Section 3.1 details the creation of individual assets, including background scenes, human motions, and texture maps. Sections 3.2 describe the rendering process and attributes of the dataset.

3.1. Dataset Assets

Background Scenes. In recent years, neural networks have significantly advanced 3D scene representation, enabling novel-view image synthesis. Notable approaches include Neural Radiance Fields (NeRF) [36], which learn a joint representation of geometry and appearance for novel-view synthesis. Another method, 3D Gaussian Splatting (3DGS) [22], represents a scene using 3D Gaussians and adopts a differentiable rasterizer for real-time rendering. These methods use multiview images to learn 3D representations, avoiding extensive capture setups. To develop our dataset, we use the 3DGS method to learn the 3D representation of the background scene.

Preliminaries: 3D Gaussian Splatting We represent the i-th Gaussian in 3DGS as:

$$G(\mathbf{p}) = o_i \, e^{-\frac{1}{2}(\mathbf{p}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{p}-\boldsymbol{\mu}_i)}, \tag{1}$$

where $\mathbf{p} \in \mathbb{R}^3$ is a xyz location, $o_i \in [0,1]$ is the opacity modeling the ratio of radiance the Gaussian absorbs, $\boldsymbol{\mu}_i \in \mathbb{R}^3$ is the center/mean of the Gaussian, and the covariance matrix $\boldsymbol{\Sigma}_i$ is parameterized by the scale $\mathbf{S}_i \in {\mathbb{R}_+}^3$ along each of the three Gaussian axes and the rotation $\mathbf{R}_i \in SO(3)$ with $\boldsymbol{\Sigma}_i = \mathbf{R}_i \mathbf{S}_i \mathbf{S}_i^\top \mathbf{R}_i^\top$. Each Gaussian is also paired with spherical harmonics [43] to model the radiance emitted towards various directions. During rendering, the 3D Gaussians are projected onto the image plane and form 2D Gaussians [62] with the covariance matrix $\boldsymbol{\Sigma}_i^{\text{2D}} = \boldsymbol{JW\Sigma}_i \boldsymbol{W}^\top \boldsymbol{J}^\top$, where $\boldsymbol{J}$ is the Jacobian of the affine approximation of the projective transformation and $\boldsymbol{W}$ is the viewing transformation. The color of a pixel is calculated via alpha blending the N Gaussians contributing to a given pixel:

$$C = \sum_{j=1}^{N} c_j \alpha_j \prod_{k=1}^{j-1} (1 - \alpha_k), \tag{2}$$

where the Gaussians are sorted from close to far, c_j is the color obtained by evaluating the spherical harmonics given viewing transform W, and α_j is calculated from the 2D Gaussian formulation (with the covariance $\boldsymbol{\Sigma}_j^{2D}$) multiplied by its opacity o_j.

We collected RGB videos from the large-scale open-source DL3DV dataset [31], which contains over 10,000 videos across diverse domains, including natural and outdoor settings, educational institutions, shopping complexes, parks, hubs, cafes, and restaurants. We selected videos based on the presence of natural occlusions, such as garlands, chairs, benches, statues, cars, and bins. Our dataset captures 3D representations of approximately 40 distinct scenes, each incorporating real-world occlusions. This approach to learning 3D representation closely resembles real-world data compared to conventional methods that rely on rendering from 3D graphic asset scenes.

SMPLX body/Human Animations. We represent the human body using the SMPL-X [40] 3D human model, defined by the function $\mathcal{M}(\theta, \beta, \psi)$, where θ represents the pose parameters, β the shape parameters, and ψ the facial expression parameters. This function outputs a body mesh $\mathcal{M} \in \mathbb{R}^{10475 \times 3}$ with 10,475 vertices. We sample approximately 400 SMPL-X 3D human motion models from the AMASS mocap dataset [34], which contains over 11,000 motion sequences with diverse body shapes and poses. Following previous work [52], we use the pre-trained human pose prior model VPoser [40] to assess pose difficulty. VPoser is a Variational Auto-Encoder model that evaluates a pose θ to be challenging if its embeddings ϵ_θ have larger norm, i.e. $||\epsilon_\theta||_2 > \tau$, where τ is empirically set to 40. We classify a challenging pose using VPoser to avoid simple movements such as walking, standing, jogging, etc. To prepare human animations for rendering, we use Blender graphics software to efficiently bake complex geometric animations over time into the Alembic format for subsequent use in rendering engines.

We ensured that motion sequences contained at least 180 frames to maintain a minimum animation duration of six seconds at 30 fps. For sequences exceeding 400 frames, we discarded the first 100 frames to remove static poses at the start. To ensure that human motion remains consistently within the occluded scene, we implemented boundary constraints that stop the motion if it moves beyond these limits. These conditions prevent humans from moving too far from the occlusion area. Additionally, we applied random rotations and translations to each sequence. We store the applied transformations to further calculate the effective global orientation of the human body with respect to the world coordinates.

Human Textures. We sourced human body texture scans from the open-source dataset provided by the SMPLitex method [5], which estimates and manipulates the full 3D appearance of humans captured from a single image. The SMPLitex dataset provides a diverse range of human texture scans, covering various skin tones, clothing styles, and genders. To ensure sample diversity in VOccl3D, we select approximately 200 distinct texture scans. Figure 4 illustrates the different clothing textures applied to the SMPL-X human model in our dataset.

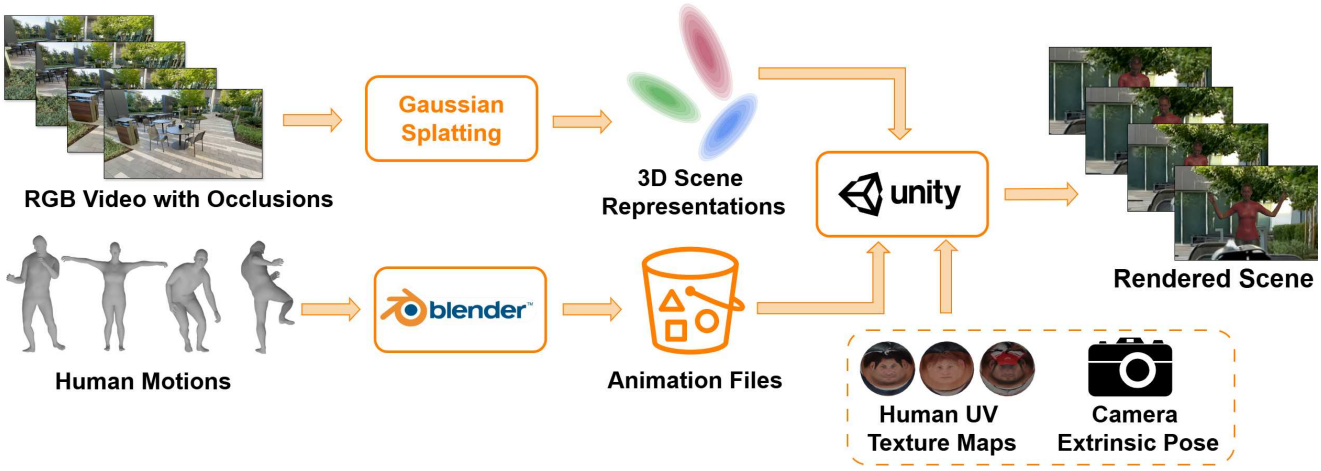

Figure 3. **An overview of our proposed dataset creation pipeline.** We generate 3D representations of natural scenes with real occlusions using 3D Gaussian Splatting [22]. Human motion sequences from the AMASS MoCap dataset [34] are processed in Blender to generate baked animation files. The generated scene representations, animation files, human texture maps, and camera extrinsic parameters are imported into the Unity rendering engine to generate video sequences of human motion under occlusion.

Figure 4. **Diversity of human texture in VOccl3D.** We use approximately 200 human texture maps from the SMPLitex dataset [5], encompassing diverse clothing styles, genders, skin tones, and ethnicities.

3.2. Dataset Rendering and Attributes

Rendering As shown in Figure 3, we used Unity Engine to render synthetic scenes by integrating background Gaussian splats, human animations, and body texture assets. We rendered video sequences at 30 fps with a resolution of 720×720, varying the camera field of view between 25 and 50 across different scenes. For each sequence, we stored ground-truth camera extrinsic and intrinsic parameters and saved the rendered RGB images in lossless PNG format. We positioned the camera to capture occluded views within the scene and applied motion constraints to human animation assets to maintain consistent occlusion throughout each sequence. To achieve realistic lighting, we employed Unity's built-in DirectionalLight asset to simulate natural illumination on the human mesh.

Dataset Attributes. Our proposed dataset, VOccl3D, includes over 250,000 images and 400 video sequences with a total runtime exceeding 2 hours and 30 minutes. The dataset features 40 background scenes with various occlusions, such as cars, garlands, benches, chairs, bins, and trees. Each scene contains 10 video sequences with variations in human motions and clothing textures. The rendered sequence exhibit diversity in body shapes, skin tones, and camera poses. We provide ground-truth annotations, including camera extrinsic and intrinsic parameters, pose, shape, global orientation, translation, gender, 2D keypoints, and a binary occlusion label for each keypoint.

Different Modalities. In addition to 3D pose and shape annotations, VOccl3D provides annotations for multiple modalities, including human silhouettes, body-part segmentation, 2D keypoint estimation, and human bounding box detection under occlusion. Researchers can use our dataset to train and evaluate methods designed to handle occlusion across these modalities.

4. Experiments and Results

In this section, we highlight the need and significance of our proposed VOccl3D dataset. We fine-tune the state-of-the-art pose estimation methods using our dataset and report the qualitative and quantitative results. Our results demonstrate improved performance on the HPS estimation task (Section 4.1) using both our proposed synthetic dataset and real-world datasets with occlusions. In Section 4.2, we show the enhanced performance of the human object detector and its impact on HPS estimation under occlusion.

4.1. Human Pose and Shape Estimation

Dataset and Implementation Setup. We use approximately 200k and 50k images for training and testing respectively in our proposed VOccl3D dataset. We fine-tune

Method	Hard-Occlusion			Medium-Occlusion			Low-Occlusion		
	MPJPE	PA-MPJPE	PVE	MPJPE	PA-MPJPE	PVE	MPJPE	PA-MPJPE	PVE
CLIFF [28]	192.22	114.35	247.41	121.70	78.56	158.46	98.82	67.64	126.92
BEDLAM-HMR [4]	167.35	102.92	214.39	102.55	68.13	134.59	86.55	54.48	110.15
BEDLAM-CLIFF [4]	154.86	99.53	199.95	90.97	65.03	119.63	74.95	52.65	96.60
HMR2.0 [13]	169.71	100.49	215.17	113.88	71.78	145.62	88.53	59.08	114.39
STRIDE with BEDLAM-CLIFF [27]	155.64	100.44	-	91.14	65.38	-	75.02	53.21	-
WHAM [48]	152.15	102.14	177.07	110.97	76.81	127.45	93.90	66.68	106.51
VOccl3D-B-CLIFF	**136.34**	**89.94**	**175.92**	82.48	**58.78**	**106.84**	**69.46**	**46.32**	**88.19**
STRIDE with VOccl3D-B-CLIFF	136.43	90.28	-	**82.37**	58.98	-	69.65	46.86	-

Table 1. **3D HPS estimation results on the test-split of VOccl3D.** The results show that VOccl3D-B-CLIFF and STRIDE, when using pseudo-labels from VOccl3D-B-CLIFF, significantly outperform other image- and video-based HPS estimation methods across hard, medium, and low occlusion categories. As expected, all HPS estimation methods exhibit a decline in performance as occlusion severity increases from low to medium to high. For evaluation, we use ground-truth bounding boxes. The best results are in **bold**.

Method	3DPW			OcclType1-3DPW			OcclType2-3DPW		
	MPJPE	PA-MPJPE	PVE	MPJPE	PA-MPJPE	PVE	MPJPE	PA-MPJPE	PVE
CLIFF [28]	73.9	46.4	87.6	98.15	62.27	118.47	99.49	62.16	119.82
BEDLAM-HMR [4]	79.0	47.6	93.1	108.62	66.53	128.05	106.19	64.05	125.66
BEDLAM-CLIFF [4]	72.0	46.6	85.0	98.71	64.26	117.41	96.80	61.32	115.33
HMR2.0 [13]	81.2	54.3	143.7	103.40	69.66	164.55	99.01	66.17	158.79
VOccl3D-B-CLIFF	72.0	47.3	84.5	95.89	63.43	**114.28**	94.36	60.44	**112.01**
VOccl3D-CLIFF	**71.10**	**45.98**	**84.25**	**95.17**	**61.83**	114.95	**93.74**	**59.66**	112.59

Table 2. **3D HPS estimation results on 3DPW, OcclType1-3DPW, and OcclType2-3DPW**. Since 3DPW is a real-world dataset with minimal occlusions, VOccl3D-CLIFF outperforms all other methods but shows only a marginal improvement over BEDLAM-CLIFF. However, on OcclType1-3DPW and OcclType2-3DPW, which contain significant occlusions on real-world dataset, our method demonstrates a notable improvement over existing approaches. The best results are noted in **bold**.

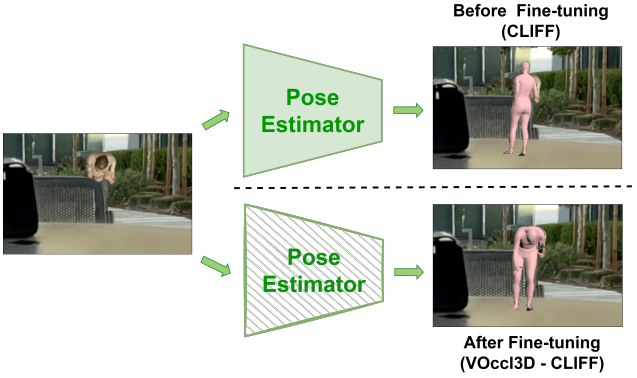

Figure 5. **Fine-tuning an off-the-shelf pose estimator.** We fine-tune a CLIFF [28] pose estimation model using our proposed VOccl3D dataset. We observe a significant improvement in the estimated mesh when using a fine-tuned CLIFF model using our dataset. Note: The pose estimator can be any off-the-shelf pose estimator, here we show results with the CLIFF model [28].

two versions of state-of-the-art model, CLIFF (trained on Human3.6M [18], MPI-INF-3DHP [35], and COCO [30] dataset) and BEDLAM-CLIFF (trained on BEDLAM and AGORA datasets) using our rendered VOccl3D dataset, resulting in VOccl3D-CLIFF and VOccl3D-B-CLIFF, respectively (refer to Figure 5). To benchmark performance, we evaluate on the test split of VOccl3D, comparing against multiple image and video based baselines, including CLIFF [28], BEDLAM-HMR [4], BEDLAM-CLIFF [4], HMR2.0 [13], STRIDE [27], and WHAM [48]. We report the mean performance by evaluating the model using a five-fold cross-validation on our dataset. To quantify occlusion levels, we annotate each 3D joint with a binary occlusion label, marking it as occluded if its corresponding 2D keypoint lies within the ground-truth segmentation mask of the visible human region. Based on the number of visible keypoints (out of 22 total), we categorize images into three occlusion levels: hard occlusion (4-9 visible keypoints), medium occlusion (10-15 visible keypoints), and low occlusion (16-20 visible keypoints).

Beyond evaluations on synthetic datasets, we assess the performance of VOccl3D fine-tuned models on real-world datasets, including 3DPW [50] and OCMotion [15]. It is noteworthy that these datasets do not contain significant occlusions. Hence, we introduce two occlusion-augmented variants of 3DPW to evaluate robustness under severe occlusions. In OcclType1-3DPW, we overlay a black patch on a randomly selected keypoint, while in OcclType2-3DPW, we occlude two random keypoints. Further details on oc-

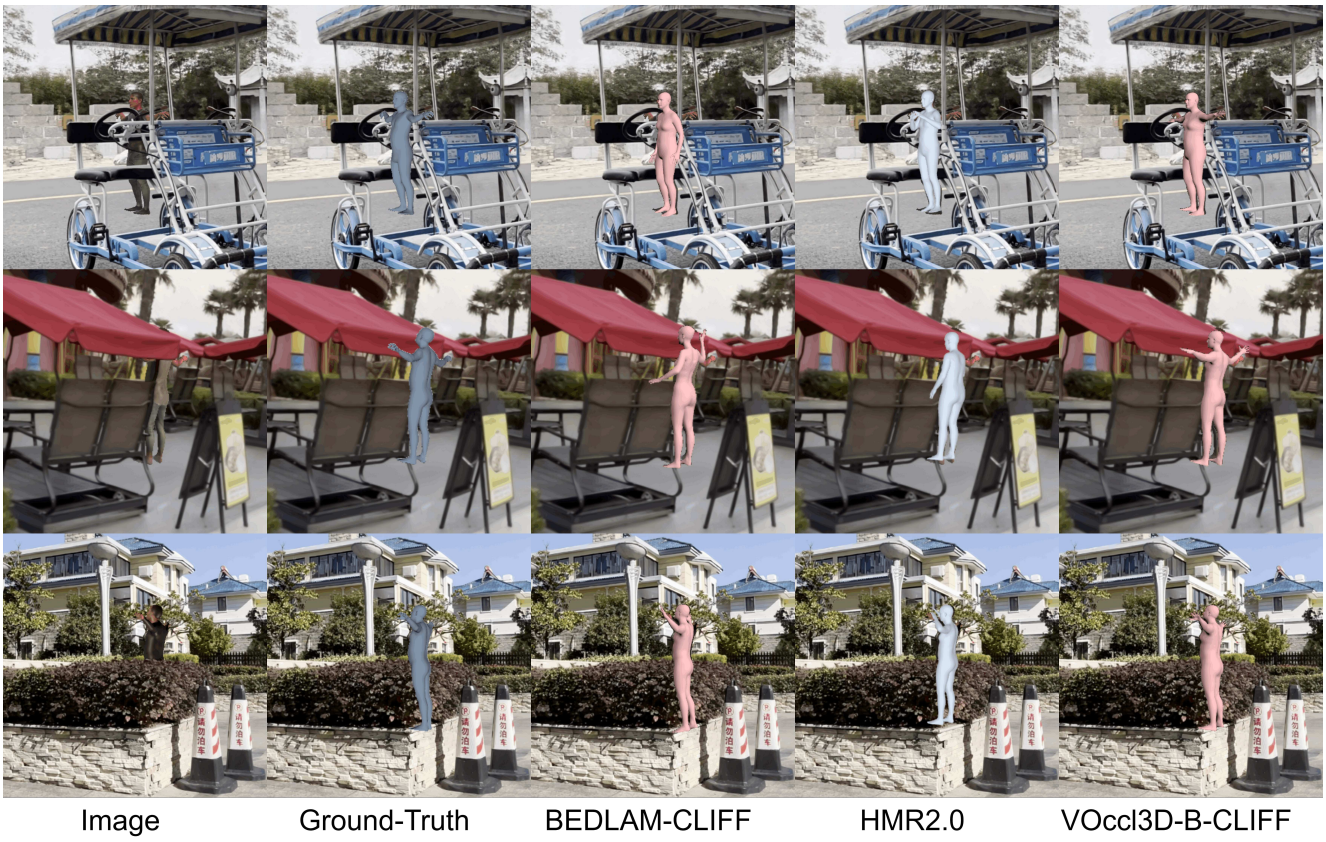

Figure 6. **Qualitative comparison of HPS estimation methods on VOccl3D dataset.** The first and second column shows RGB image and ground-truth human mesh. Column third and fourth compare HPS estimation using the BEDLAM-CLIFF [4] and HMR2.0 [13] methods. The final column (VOccl3D-B-CLIFF) presents results obtained by fine-tuning the CLIFF model on the VOccl3D dataset. These results demonstrate superior performance, particularly in scenarios with heavy occlusion.

clusion variants of 3DPW and implementation details are provided in the supplementary material.

Results. We present benchmarking results on the test split of our proposed VOccl3D dataset in Table 1. Our fine-tuned VOccl3D-B-CLIFF significantly outperforms existing state-of-the-art image and video-based methods across all occlusion categories. Notably, the performance of all previous methods degrades as occlusion severity increases, highlighting the inherent challenges of HPS estimation under occlusions. We also compare our approach with the plug-and-play method STRIDE [27], which performs 3D pose estimation and reports only MPJPE and PA-MPJPE errors. Our results show that STRIDE achieves superior performance when leveraging pseudo-labels from the VOccl3D-B-CLIFF model compared to those from the original BEDLAM-CLIFF model, further validating the effectiveness of our dataset. In Table 2 and Table 3, we report performance on the 3DPW [49] and OCMotion [16] datasets, respectively. We observe that both VOccl3D-B-CLIFF and VOccl3D-CLIFF outperform all the previous methods, demonstrating their effectiveness in real-world scenarios. However, since these datasets contain minimal occlusions, the improvements over CLIFF remain marginal. To further assess robustness under heavy occlusion, we evaluate on OcclType1-3DPW and OcclType2-3DPW, real-world datasets with significant occlusions. Table 2 demonstrates that our fine-tuned model achieves substantial improvements, underscoring its robustness in occluded environments. Table 3 presents results on the OCMotion dataset, which features lower levels of occlusion. We can observe that VOccl3D-CLIFF outperforms all comparison methods and demonstrates performance comparable to CLIFF [28]. This is expected, as both models are pre-trained on large-scale real-world datasets captured in controlled lab environments similar to OCMotion. Figure 6 presents the qualitative results on VOccl3Ddataset, showcasing the improvement of our fine-tuned model against other state-of-the-art HPS estimation methods. Additional results are provided in supplementary material.

4.2. Impact of Human Detector on HPS Estimation

Dataset and Implementation Setup. Human object detectors play a crucial role in HPS estimation tasks, particularly

Method	OCMotion		
	MPJPE	PA-MPJPE	PVE
CLIFF [28]	**64.15**	40.16	79.80
BEDLAM-HMR [4]	73.96	41.94	92.70
BEDLAM-CLIFF [4]	66.80	41.79	83.86
HMR2.0 [13]	69.94	43.00	87.07
VOccl3D-B-CLIFF	65.96	40.59	81.96
VOccl3D-CLIFF	64.29	**39.64**	**78.56**

Table 3. **Qualitative results of HPS estimation on OCMotion.** Our results show that VOccl3D-CLIFF outperforms all HPS estimation methods except CLIFF [28]. Both CLIFF and VOccl3D-CLIFF exhibit comparable performance, as both are pre-trained on real-world datasets captured in controlled lab environments, similar to OCMotion [16]. The best results are in **bold**.

Method	3DPW		OCMotion	
	mAP50	mAP75	mAP50	mAP75
YOLO11	58.99	47.14	98.84	91.80
VOccl3D-YOLO11	**59.89**	**48.26**	**99.10**	**91.95**

Table 4. **Quantitative results of YOLO11 on 3DPW and OCMotion datasets.** The first row presents the human object detection performance of the pre-trained YOLO11 model, while the second row shows the results after fine-tuning with the VOccl3D dataset. The best results are in **bold**. We observe an improvement in detection accuracy across both datasets after fine-tuning.

when inferring from in-the-wild RGB images [4, 27, 28]. Ideally, HPS estimation performs optimally with ground-truth bounding boxes; however, most human object detectors struggle significantly under occlusion. To address this, we fine-tune the recent YOLO11 detector [19] on the combined training split of VOccl3D and MS COCO, referring to the fine-tuned model as VOccl3D-YOLO11. We show evaluations on 3DPW [49] and OCMotion [16] datasets.

Results. Table 4 compares the detection performance of the pre-trained YOLO11 and VOccl3D-YOLO11 models on the real-world 3DPW [49] and OCMotion [16] datasets. Since OCMotion is a single-human dataset with minimal occlusions, the detector achieves high mAP scores. This is likely because most body parts remain visible, making box detection easier. However, due to the presence of multiple individuals in 3DPW, the overall mAP scores remain lower. In short, VOccl3D-YOLO11 shows significant improvement on datasets with high occlusions. We provide qualitative results in the supplementary material.

Table 5 further evaluates the impact of human detection on HPS estimation. The first, second, and third rows represent the performance of VOccl3D-CLIFF when human detections are sourced from ground truth, YOLO11, and VOccl3D-YOLO11, respectively. As expected, the best performance is achieved when using ground-truth detections. However, fine-tuning the detector with VOccl3D significantly improves HPS estimation compared to detections

Method	3DPW			OCMotion		
	MPJPE	PA-MPJPE	PVE	MPJPE	PA-MPJPE	PVE
VOccl3D-CLIFF w/GT	71.10	45.98	84.25	64.29	39.64	78.56
VOccl3D-CLIFF w/YOLO11	116.52	63.35	139.74	67.16	41.30	83.00
VOccl3D-CLIFF w/VOccl3D-YOLO11	114.85	62.66	137.19	66.65	41.15	82.40

Table 5. **3D HPS estimation results using the VOccl3D-CLIFF model with different bounding box sources.** This table presents HPS estimation performance on the 3DPW and OCMotion datasets. The first row reports results using ground-truth bounding boxes, achieving the highest accuracy. The second and third rows show performance when detections are obtained from the pre-trained YOLO11 and fine-tuned VOccl3D-YOLO11 models, respectively. Notably, the third row demonstrates a significant improvement over the second, highlighting the impact of detections on HPS estimation.

from the pre-trained YOLO11. These results underscore the critical role of human detection quality in enhancing HPS performance under occlusion.

Our experiments show that fine-tuning an existing HPS estimation model with a large synthetic dataset containing occlusions enhances its performance on both real and synthetic datasets. We observe significant improvements over baseline methods, particularly in scenarios with heavy occlusions. Additionally we demonstrate that the poor results are not solely due to HPS estimations errors but also stem from failures in bounding box detection. The VOccl3D dataset proves effective in refining pose estimation methods and bounding box predictions under occlusions.

5. Conclusion

We introduce VOccl3D, a novel large-scale video dataset with synthetic humans in real-world scenes under occlusions for 3D human pose and shape estimation. By leveraging a rendering-based approach, VOccl3D eliminates the need for costly and labor-intensive data collection while providing a diverse and realistic dataset for occlusion-aware research. Through extensive qualitative and quantitative evaluations, we demonstrate that fine-tuning existing HPS estimation methods with VOccl3D significantly enhances their performance on real-world datasets with occlusions and on the test split of VOccl3D dataset. Furthermore, we improve the human object detector, YOLO11 under occluded conditions using VOccl3D, highlighting the impact of detections in achieving robust HPS estimation in occluded scenarios. Beyond HPS estimation, VOccl3D serves as a comprehensive benchmark for evaluating methods across multiple occlusion-aware tasks, including human body-part segmentation, 2D/3D pose estimation, and human bounding box detection. Hence, VOccl3D sets a new standard for occluded human benchmarks, offering a valuable dataset for advancing occlusion-robust research.

Acknowledgment

This work was partially supported by USDA NRI grant 2021-67022-33453, NSF grants CMMI-2326309 and CNS-2312395. This research was also partially supported by the Office of the Director of National Intelligence (ODNI), specifically through the Intelligence Advanced Research Projects Activity (IARPA), under contract number [2022-21102100007]. The views and conclusions in this research reflect those of the authors and should not be construed as officially representing the policies, whether explicitly or implicitly, of ODNI, IARPA, or the U.S. Government. Nevertheless, the U.S. Government retains the authorization to reproduce and distribute reprints for official government purposes, regardless of any copyright notices included.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 12

[2] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[3] Peter Bauer, Arij Bouazizi, Ulrich Kressel, and Fabian B. Flohr. Weakly supervised multi-modal 3d human body pose estimation for autonomous driving. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–7, 2023. 2

[4] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion, 2023. 2, 3, 6, 7, 8, 12, 14

[5] Dan Casas and Marc Comino-Trinidad. SMPLitex: A Generative Model and Dataset for 3D Human Texture Estimation from Single Image. In *British Machine Vision Conference (BMVC)*, 2023. 4, 5

[6] Yu Cheng, Bo Yang, Bo Wang, Yan Wending, and Robby Tan. Occlusion-aware networks for 3d human pose estimation in video. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 723–732, 2019. 3

[7] Henry M Clever, Zackory Erickson, Ariel Kapusta, Greg Turk, Karen Liu, and Charles C Kemp. Bodies at rest: 3d human pose and shape estimation from a pressure image using synthetic data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6215–6224, 2020. 3

[8] Mickael Cormier, Aris Clepe, Andreas Specker, and Jürgen Beyerer. Where are we with human pose estimation in real-world surveillance? In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 591–601, 2022. 2

[9] R James Cotton. Posepipe: Open-source human pose estimation pipeline for rehabilitation research. *Archives of Physical Medicine and Rehabilitation*, 103(12):e161–e162, 2022. 2

[10] Yufeng Cui and Yimei Kang. Multi-modal gait recognition via effective spatial-temporal feature fusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17949–17957, 2023. 2

[11] Yang Fu, Shibei Meng, Saihui Hou, Xuecai Hu, and Yongzhen Huang. Gpgait: Generalized pose-based gait recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19538–19547, 2023. 2

[12] Lei Geng, Wenzhu Yang, Yanyan Jiao, Shuang Zeng, and Xinting Chen. A multilayer human motion prediction perceptron by aggregating repetitive motion. *Mach. Vision Appl.*, 34(6), 2023. 13

[13] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023. 6, 7, 8, 14

[14] Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafiirah Hosenie, Thomas J Cashman, Julien Valentin, Darren Cosker, et al. Look ma, no markers: Holistic performance capture without the hassle. *ACM Transactions on Graphics*, 43(6): 1–12, 2024. 2, 3

[15] Buzhen Huang, Tianshu Zhang, and Yangang Wang. Object-occluded human shape and pose estimation with probabilistic latent consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5010–5026, 2022. 6

[16] Buzhen Huang, Tianshu Zhang, and Yangang Wang. Object-occluded human shape and pose estimation with probabilistic latent consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5010–5026, 2023. 2, 7, 8

[17] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 12

[18] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 2, 3, 6

[19] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. 3, 8, 12

[20] Sardor Juraev, Akash Ghimire, Jumabek Alikhanov, Vijay Kakani, and Hakil Kim. Exploring human pose estimation and the usage of synthetic data for elderly fall detection in real-world surveillance. *IEEE Access*, 10:94249–94261, 2022. 2

[21] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2017. 2, 3

[22] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 2, 4, 5

[23] Do-Yeop Kim and Ju-Yong Chang. Attention-based 3d human pose sequence refinement network. *Sensors*, 21(13), 2021. 3

[24] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 13

[25] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021. 2

[26] Farnoosh Koleini, Muhammad Usama Saleem, Pu Wang, Hongfei Xue, Ahmed Helmy, and Abbey Fenwick. Biopose: Biomechanically-accurate 3d pose estimation from monocular videos. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6330–6339. IEEE, 2025. 2

[27] Rohit Lal, Saketh Bachu, Yash Garg, Arindam Dutta, Calvin-Khang Ta, Dripta S. Raychaudhuri, Hannah Dela Cruz, M. Salman Asif, and Amit K. Roy-Chowdhury. Stride: Single-video based temporally continuous occlusion robust 3d pose estimation, 2024. 2, 3, 6, 7, 8

[28] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation, 2022. 2, 3, 6, 7, 8, 12, 14

[29] Kevin Lin, Chung-Ching Lin, Lin Liang, Zicheng Liu, and Lijuan Wang. Mpt: Mesh pre-training with transformers for human pose and mesh reconstruction. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3403–3413, 2022. 13

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 6, 12

[31] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 4

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2, 3

[33] Zhengyi Luo, S. Alireza Golestaneh, and Kris M. Kitani. 3d human motion estimation via motion compression and refinement. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. 2, 3

[34] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 4, 5, 12

[35] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 International Conference on 3D Vision (3DV)*, pages 506–516, 2017. 6, 12

[36] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 4

[37] Sara Moccia, Lucia Migliorelli, Virgilio Carnielli, and Emanuele Frontoni. Preterm infants' pose estimation with spatio-temporal features. *IEEE Transactions on Biomedical Engineering*, 67(8):2370–2380, 2019. 2

[38] Aron Monszpart, Paul Guerrero, Duygu Ceylan, Ersin Yumer, and Niloy J. Mitra. imapper: interaction-guided scene mapping from monocular videos. *ACM Trans. Graph.*, 38(4), 2019. 2

[39] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora: Avatars in geography optimized for regression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13468–13478, 2021. 2, 12

[40] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 4

[41] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7745–7754, 2018. 3

[42] Miroslav Purkrábek and Jiří Matas. Improving 2d human pose estimation across unseen camera views with synthetic data. *arXiv preprint arXiv:2307.06737*, 2023. 3

[43] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 497–500, 2001. 4

[44] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[45] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, page 1530–1538. JMLR.org, 2015. 3

[46] Tim Salzmann, Hao-Tien Lewis Chiang, Markus Ryll, Dorsa Sadigh, Carolina Parada, and Alex Bewley. Robots that can see: Leveraging human pose for trajectory prediction. *IEEE Robotics and Automation Letters*, 8(11):7090–7097, 2023. 2

[47] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[48] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2070–2080, 2023. 2, 3, 6

[49] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, 2018. 2, 7, 8, 13

[50] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 6

[51] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3d human pose estimation with normalizing flows. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[52] Zhenzhen Weng, Laura Bravo-Sánchez, and Serena Yeung-Levy. Diffusion-hpc: Synthetic data generation for human mesh recovery in challenging domains. In *2024 International Conference on 3D Vision (3DV)*, pages 257–267. IEEE, 2024. 4

[53] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, et al. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20282–20292, 2023. 3

[54] Frank Yu, Mathieu Salzmann, Pascal Fua, and Helge Rhodin. Pcls: Geometry-aware neural reconstruction of 3d pose with perspective crop layers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9060–9069, 2021. 3

[55] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[56] Andrei Zanfir, Mihai Zanfir, Alex Gorban, Jingwei Ji, Yin Zhou, Dragomir Anguelov, and Cristian Sminchisescu. HUM3DIL: Semi-supervised multi-modal 3d humanpose estimation for autonomous driving. In *6th Annual Conference on Robot Learning*, 2022. 2

[57] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: A plug-and-play network for refining human poses in videos. In *European Conference on Computer Vision*. Springer, 2022. 3

[58] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2020. 2

[59] Yi Zhang, Pengliang Ji, Adam Kortylewski, Angtian Wang, Jieru Mei, and Alan L Yuille. 3D-Aware Neural Body Fitting for Occlusion Robust 3D Human Pose Estimation. In *The IEEE/CVF International Conference on Computer Vision*, 2023. 2

[60] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G. Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4966–4975, 2015. 3

[61] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 13

[62] Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. Surface splatting. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 371–378, 2001. 4

Supplementary Material

A. Implementation details.

A.1. Human Pose and Shape estimation.

We fine-tune CLIFF [28] and BEDLAM-CLIFF [4] for HPS estimation using approximately 200k images from our VOccl3D dataset. CLIFF is trained on real 2D datasets such as COCO [30] and MPII [1], as well as 3D datasets like Human3.6M [17] and 3DHP [35], while BEDLAM-CLIFF is originally trained on synthetic datasets such as BEDLAM [4] and AGORA [39]. We fine-tune these models on a single NVIDIA GeForce RTX 3090 Ti GPU. We adopt hyperparameters and loss functions from [4] for fine-tuning. We optimize the models using the Adam optimizer with a learning rate of 0.00005 and zero weight decay. To prevent overfitting, we employ early stopping. We use a batch size of 64 and resize input images to 224×224 dimension.

We report errors after converting SMPL-X bodies to SMPL using a pre-trained joint regressor mapping and aligning the pelvis of these bodies. We evaluate CLIFF, BEDLAM-CLIFF, BEDLAM-HMR, HMR2.0, WHAM, and STRIDE by re-running their evaluations using the official code repositories.

We create two variants of the 3DPW dataset, OcclType1-3DPW and OcclType2-3DPW, by overlaying black patches to evaluate performance on highly occluded real-world datasets. OcclType1-3DPW is generated by randomly adding a black patch over a single 2D keypoint from the 22 openpose joints, while OcclType2-3DPW contains images with two black patches placed on random 2D keypoints. The added patches are square-shaped, with dimensions covering 60% of the human height in OcclType1-3DPW and 40% of the human height in OcclType2-3DPW. Figure 8 illustrates sample images from OcclType1-3DPW and OcclType2-3DPW. We follow the same evaluation procedure for real-world datasets, including 3DPW, OcclType1-3DPW, OcclType2-3DPW, and OCMotion, as we do for the VOccl3D dataset.

Evaluation metrics. Following prior works, we use standard metrics to report the performance of human pose and shape estimation. MPJPE and PVE represent the average error in joints and vertices respectively after aligning the pelvis. PA-MPJPE reports the average error after aligning the rotation and scale. All errors are in mm.

A.2. Human detector.

We conduct our experiments on the YOLO11 detector using the official Ultralytics codebase [19]. The original YOLO11 model is pre-trained on the MS COCO dataset [30]. To enhance its performance under occlusions, we fine-tune YOLO11 on the combined train split of VOccl3D and MS COCO, resulting in VOccl3D-YOLO11. We fine-tune the model for 50 epochs with a batch size of 32 on a single NVIDIA GeForce RTX 3090 Ti GPU. Following [19], we resize input images to 640×640 and train using a learning rate of 0.01 with a weight decay of 0.0005. Additionally, we set the loss function weights to 7.5 for the bounding box component and 0.5 for the classification component to optimize detection performance.

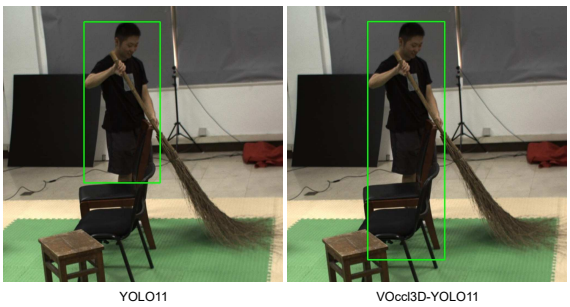

Figure 7. **Human detection under occlusion on OCMotion using YOLO11.** The left image illustrates detection performance with the pre-trained YOLO11, while the right image shows improved detection after fine-tuning YOLO11 with the VOccl3D dataset, resulting in VOccl3D-YOLO11.

Figure 7 shows the qualitative performance of YOLO11 and VOccl3D-YOLO11, where we show an improved performance of VOccl3D-YOLO11 under high occlusions.

Evaluation metrics. We evaluate detector performance using mean Average Precision (mAP) at Intersection over Union (IoU) thresholds of 0.50 and 0.75, referred to as mAP50 and mAP75, respectively. Unlike standard bounding box labels that include only visible human regions, we provide bounding box annotations that cover the entire human body, including both visible and occluded parts.

B. Additional related works.

Datasets for Pose Estimation Previous works have proposed several datasets for HPSE, which are either video-based or image-based. One of the pioneers in this field is the CMU Motion Capture dataset which primarily contained 3D skeletal data without RGB images. This dataset included a wide range of activities like dancing, walking, and sports and served as a cornerstone for tasks like animation, pose estimation, and gaming. Further, in 2016, the MSCOCO dataset [30] was released which initially contained over 200,000 labeled images covering 80 object categories, including humans. The scale of this dataset provided a wealth of data that was unprecedented for pose estimation tasks at the time. Additionally, MSCOCO introduced keypoint annotations for human pose estimation, providing 17 key points per person. The Archive of Motion Capture As Surface Shapes (AMASS) dataset [34], introduced in [34], is a large human motion database that unifies various opti-

Figure 8. Samples of OcclType1-3DPW (top row) and OcclType2-3DPW (bottom row) dataset.

cal marker-based motion capture datasets under a common framework and parameterization. This dataset contains 40 hours of human motion data, spanning over 300 subjects, and motivated large-scale pre-training in a variety of follow-up HPS works [12, 24, 29, 61]. The recent 3D Poses in the Wild (3DPW) dataset [49] is a widely-used benchmark for evaluating 3D human pose estimation methods in natural, unstructured environments, providing accurate 3D pose annotations derived from synchronized video and inertial measurement unit (IMU) data. This dataset comprises over 51,000 frames and across 60 video sequences. Although these datasets fueled the state-of-the-art methods but contain limited occlusions in their samples. This makes methods trained on these datasets vulnerable to occlusions, limiting their ability to generalize to unseen scenarios with significant occlusions.

C. Qualitative examples

In this section, we present the qualitative results of our fine-tuned model, VOccl3D-B-CLIFF, in comparison with other HPS estimation methods. Figure 9 illustrates qualitative results on the OcclType2-3DPW dataset, while Figure 10 provides additional qualitative comparisons on the test split of VOccl3D. We observe the superior performance of VOccl3D-B-CLIFF across multiple datasets. Additionally, Figure 11 showcases further sample images from the VOccl3D dataset.

D. Limitations and Future Work

Our work highlights the need and importance of a large-scale, realistic occluded human dataset for performing the task of human pose and shape estimation. By releasing this dataset and the associated tools for repopulation, we aim to enable the research community to systematically evaluate their algorithms under challenging occlusion scenarios.

Currently, the visual quality of our synthetic humans is limited by the lack of open-source high-fidelity assets, such as garments, hairstyles, footwear, and diverse human motions, which are constrained by the AMASS dataset. Moreover, our rendering pipeline relies on predefined camera poses to generate images with substantial occlusions. A promising direction for future work would be to develop an end-to-end framework that can automatically generate occlusion-rich sequences without requiring externally provided camera parameters.

Although the VOccl3D dataset offers realistic occlusion scenarios, a noticeable gap remains between synthetic and real-world data. Bridging this sim-to-real gap represents an important avenue for future research in realistic human pose estimation. Additionally, our dataset holds potential utility for broader research efforts focused on occlusion-aware learning across various modalities, including human silhouette extraction, body-part segmentation, 2D keypoint estimation, and bounding box detection

Figure 9. **Qualitative comparison of HPS estimation methods on OcclType2-3DPW dataset.** Column 1 represents input RGB image. Columns 2–4 compare HPS estimation using the CLIFF [28], BEDLAM-CLIFF [4], and HMR2.0 [13] methods. The final column (VOccl3D-B-CLIFF) presents results obtained by fine-tuning the CLIFF model on the VOccl3D dataset.

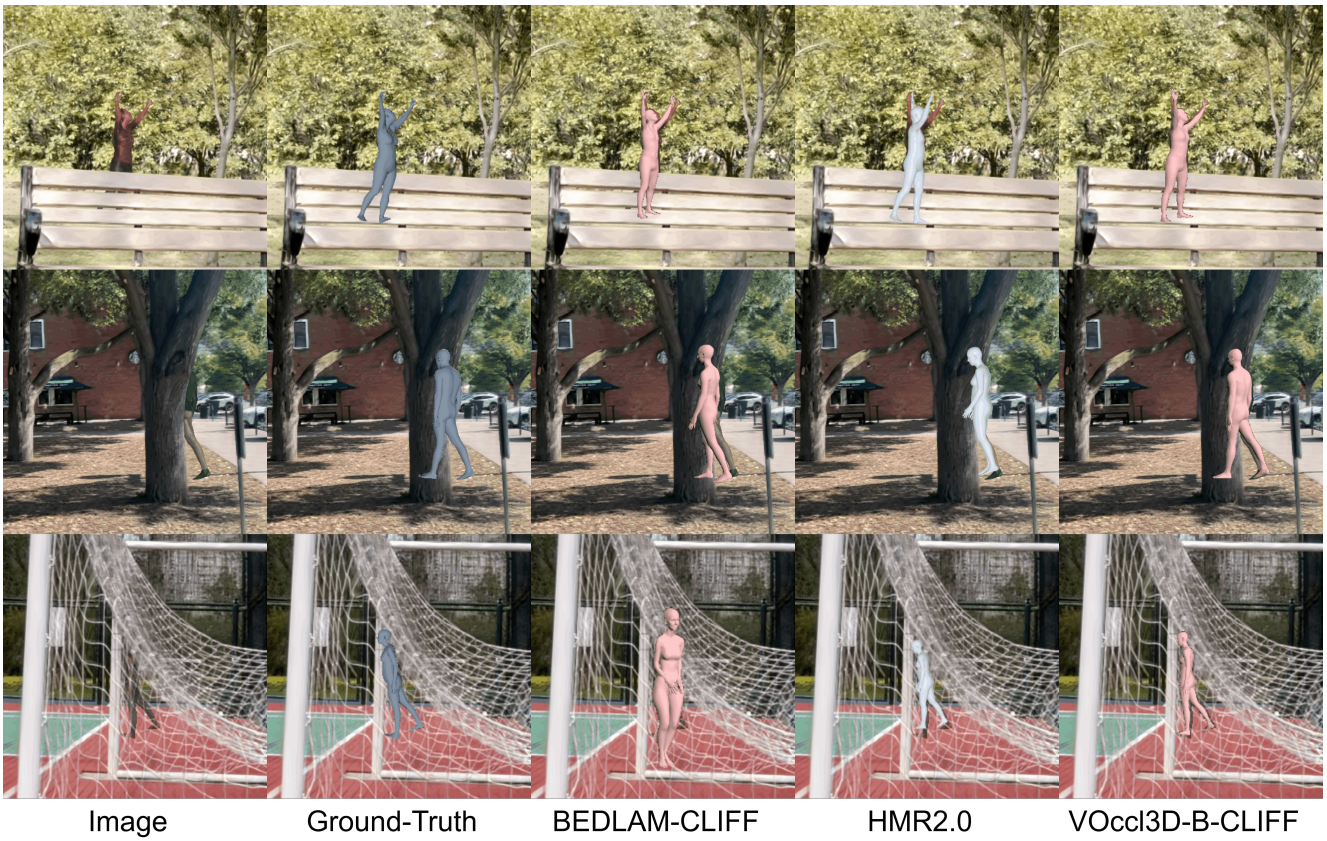

Figure 10. **Qualitative comparison of HPS estimation methods on VOccl3D dataset.** Column 1 and 2 represents input RGB image and ground truth pose. Columns 3 and 4 compare HPS estimation using the BEDLAM-CLIFF [4], and HMR2.0 [13] methods. The final column (VOccl3D-B-CLIFF) presents results obtained by fine-tuning the CLIFF model on the VOccl3D dataset.

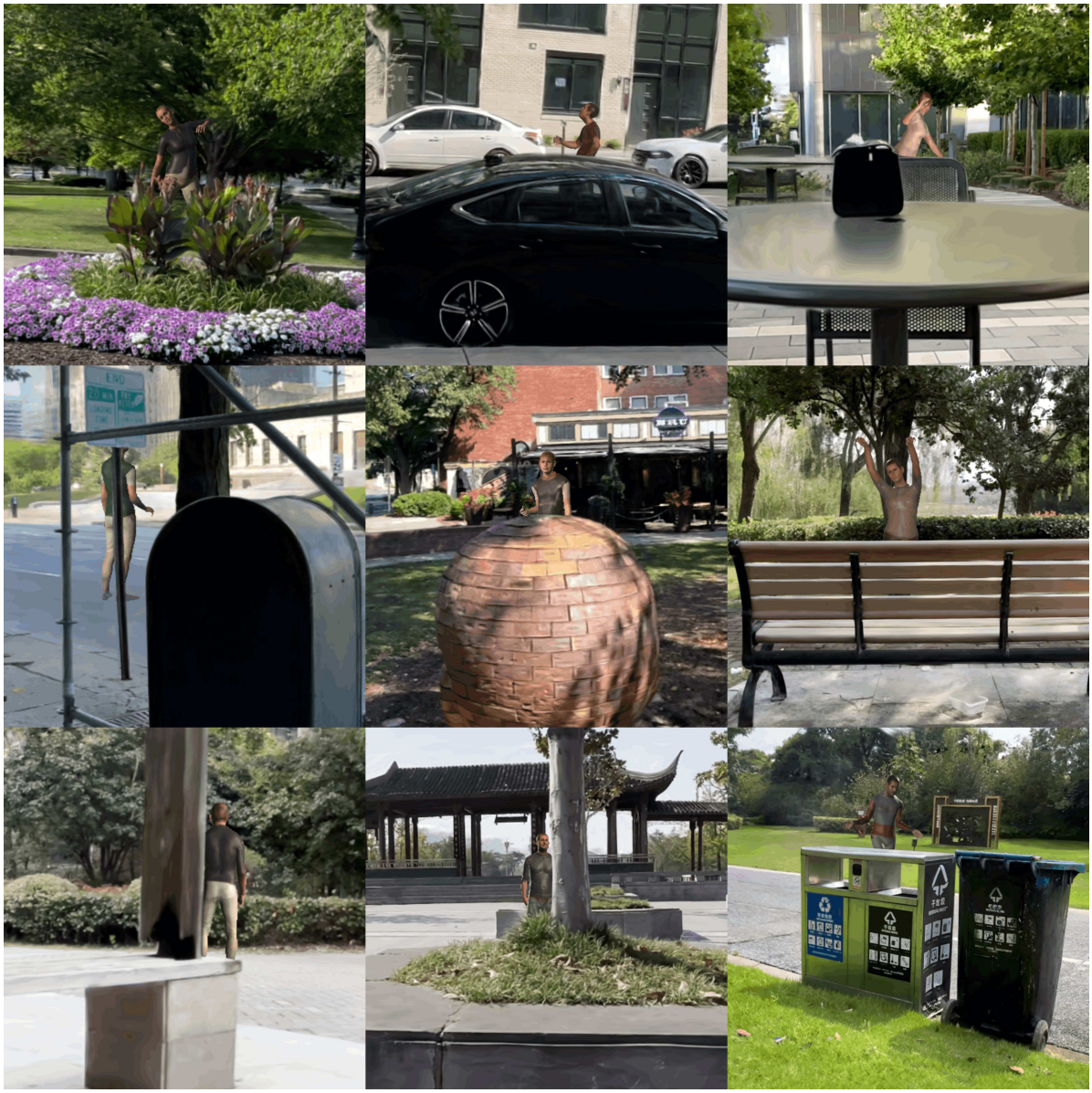

Figure 11. **Samples of VOccl3D dataset.** The samples from VOccl3D dataset illustrates various diversity in real occlusions, human motions, and clothing textures.

Datasets	#Sub	#Frames	Image	Subj/image	Motion	Ground-Truth	Occlusion	Multi-level Occlusion	Video data
SURREAL	145	∼6.5M	composite	1	>2k	SMPL	No	No	No
MPI-INF-3DHP-Train	8	>1.3M	mixed/composite	1	8+	3D joints	No	No	Yes
AGORA	>350	∼18k	rendered	5-15	n/a	SMPL-X	Yes	No	No
BEDLAM	217	380k	rendered	1-10	2311	SMPL-X	No	No	Yes
SynthMoCap	∼200	∼100k	rendered	1-4	n/a	SMPL-X	No	No	No
OCMotion	8	300k	captured	1	43	SMPL	Yes	No	Yes
VOccl3D	∼200	∼250k	rendered	1	400	SMPL-X	Yes	Yes	Yes

Table 6. Comparison of synthetic datasets and real dataset with occlusion for 3D human pose estimation.