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

Yash Garg[1] Saketh Bachu[1] Arindam Dutta[1] Rohit Lal[†1] Sarosij Bose[1]
Calvin-Khang Ta[‡1] M. Salman Asif[1] Amit Roy-Chowdhury[1]
[1]University of California, Riverside
{ygarg002,sbach008,adutt020,rlal011,sbose007,cta003,sasif,amitrc}@ucr.edu

A. Implementation details.

A.1. Human Pose and Shape estimation.

We fine-tune CLIFF [8] and BEDLAM-CLIFF [2] for HPS estimation using approximately 200k images from our VOccl3D dataset. CLIFF is trained on real 2D datasets such as COCO [10] and MPII [1], as well as 3D datasets like Human3.6M [5] and 3DHP [12], while BEDLAM-CLIFF is originally trained on synthetic datasets such as BEDLAM [2] and AGORA [13]. We fine-tune these models on a single NVIDIA GeForce RTX 3090 Ti GPU. We adopt hyperparameters and loss functions from [2] 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 2 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.

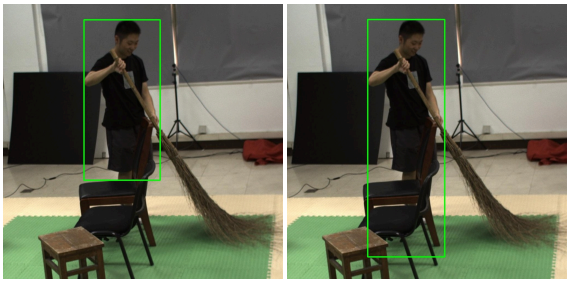

Figure 1. **Human detection under occlusion on OCMotion using YOLO11.** The left image illustrates detection performance with the pre-trained YOLO11, while the right image shows improved detection after fine-tuning YOLO11 with the VOccl3D dataset, resulting in VOccl3D-YOLO11.

A.2. Human detector.

We conduct our experiments on the YOLO11 detector using the official Ultralytics codebase [6]. The original YOLO11 model is pre-trained on the MS COCO dataset [10]. 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 [6], 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 2. Samples of OcclType1-3DPW (top row) and OcclType2-3DPW (bottom row) dataset.

Figure 1 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 [10] 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 [11], introduced in [11], is a large human motion database that unifies various optical 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 [3, 7, 9, 15]. The recent 3D Poses in the Wild (3DPW) dataset [14] 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 3 illustrates qualitative results on the OcclType2-3DPW dataset, while Figure 4 provides additional qualitative comparisons on the test split of VOccl3D. We observe the superior performance of VOccl3D-B-CLIFF across multiple datasets. Additionally, Figure 5 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

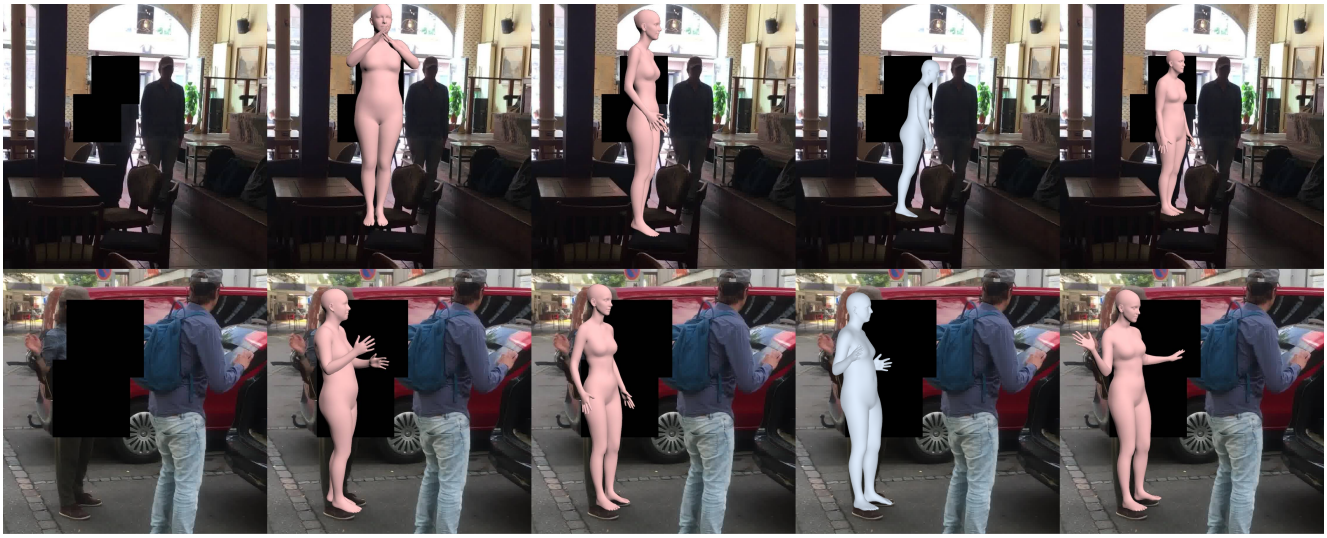

Figure 3. **Qualitative comparison of HPS estimation methods on OcclType2-3DPW dataset.** Column 1 represents input RGB image. Columns 2–4 compare HPS estimation using the CLIFF [8], BEDLAM-CLIFF [2], and HMR2.0 [4] methods. The final column (VOccl3D-B-CLIFF) presents results obtained by fine-tuning the CLIFF model on the VOccl3D dataset.

Figure 4. **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 [2], and HMR2.0 [4] methods. The final column (VOccl3D-B-CLIFF) presents results obtained by fine-tuning the CLIFF model on the VOccl3D dataset.

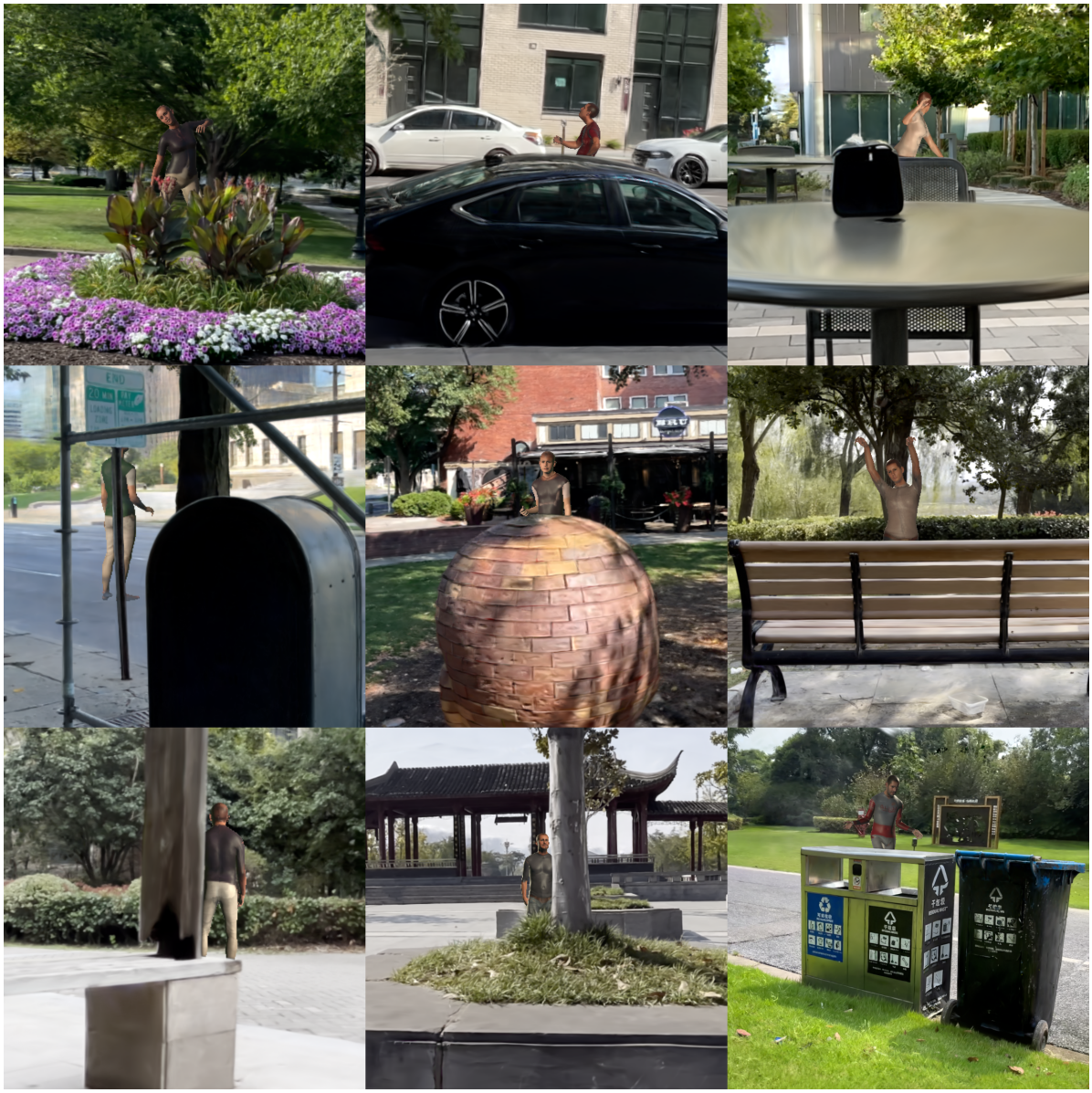

Figure 5. **Samples of VOccl3D dataset.** The samples from VOccl3D dataset illustrates various diversity in real occlusions, human motions, and clothing textures.

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

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

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. 1

[2] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion, 2023. 1, 3

[3] 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. 2

[4] 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. 3

[5] 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. 1

[6] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. 1

[7] 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

[8] 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. 1, 3

[9] 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. 2

[10] 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. 1, 2

[11] 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

[12] 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. 1

[13] 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. 1

[14] 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

[15] 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. 2