MVMP-HMR: MULTIVIEW MULTI-PERSON HUMAN MESH RECOVERY UNDER LARGE SCENES WITH OCCLUSIONS

Anonymous authors
Paper under double-blind review

Figure 1: Comparison between single-view HMR datasets and our proposed MVMP-HMR dataset. The left shows images from GTA-Human (Cai et al., 2024b), EHF (Pavlakos et al., 2019), AGORA (Patel et al., 2021), and BEDLAM (Black et al., 2023) datasets from top to bottom. Our multiview images and ground truth meshes are shown on the right, containing larger scenes and more persons. The red box indicates areas with severe occlusions.

ABSTRACT

Human mesh recovery (HMR) refers to recovering the human 3D meshes from images. Most existing HMR tasks focus on multi-person from a single image or a single person from multiple views. And the evaluation benchmarks used in these methods usually contain quite small numbers of humans or under small scenes, which is unreliable for real applications with severe occlusions. Thus, we present Multiview Multi-Person HMR (MVMP-HMR), a multiview model for multi-person whole-body human mesh recovery from multi-view images under occluded scenes. Specifically, MVMP-HMR first fuses multiple views to obtain a 3D feature volume for all persons, and then the pelvis joint from a 3D pose estimation net is utilized to acquire the human query of each person from the 3D feature volume. Finally, the human queries are cross-attentioned with the 3D feature volume and integrated to decode each person's 3D meshes. Besides, two novel losses are put forward to further enhance the model performance: the orientation loss and the 3D joint density loss, dealing with the orientation and pose ambiguities in the mesh predictions under the occluded scenes. Furthermore, a large synthetic MVMP-HMR dataset is proposed, which consists of 15 multiview scenes with up to 50 camera views and 30 persons. Experiments demonstrate that the existing state-of-the-art (SOTA) HMR methods cannot perform well on the proposed large MVMP-HMR benchmark, and the proposed MVMP-HMR model's advantages over existing SOTAs under large scenes with severe occlusions.

1 INTRODUCTION

Human mesh recovery (HMR) predicts the human 3D meshes from images or image crops, which has important applications in autonomous driving, digital games, or AR/VR, *etc*. Most existing HMR methods focus on recovering human meshes for scenes with a quite limited people number (usually < 15 in total), either with a single person from single images or multi-crops, or multi-persons from single images. Besides, the evaluation benchmarks used in the latest methods are usually under small scenes, with few occlusions (see Figure 1 left). This is not practical for real-world applications where there might be massive crowds in large scenes with severe occlusions. Thus, the existing HMR methods have not been evaluated under more complicated conditions with both larger human numbers and more severe occlusions, whose performance is not ensured.

To solve the problem and extend the HMR task to more complicated scenes, in this paper, we present MVMP-HMR (as in Figure 2), a novel model for multi-person whole-body human mesh recovery from multi-view images, which fuses multiview clues to handle the severe occlusions in large scenes with more humans. Specifically, MVMP-HMR extracts single-view features and projects them to the 3D space, and then the projected multi-view features are averaged to obtain a complete 3D feature volume for the whole scene. Besides, a 3D pose estimation branch is adopted to predict the pelvis joint location of each person, and the predicted pelvis joint is used to acquire the human queries by sampling at the locations from the previously fused 3D feature volume. Then the human queries and the 3D feature volume are both fed into the human transformer block (HTB) where both are fused via cross-attention layers. Finally, the output of HTB is decoded to regress the SMPL-X parameters.

To deal with the human orientation and pose ambiguities in the predicted SMPL-X parameters under the occluded scenes, in addition to common parameter regression losses used in single-view HMR SOTA (Baradel et al., 2024), we put forward two novel losses: the **orientation loss** and the **3D joint density loss**. The orientation loss $\mathcal{L}_{\mathcal{O}}$ is the supervision of the human mesh's orientation in the real-world coordinates. The 3D joint density loss $\mathcal{L}_{denj3d}$ supervises the 3D joints in the predicted human mesh via 3D joint density maps instead of direct joint coordinate regression. Both provide stronger supervision in the 3D space and handle the orientation and pose ambiguities in the MVMP-HMR task better, further enhancing the model performance (see results in Sec. 4.4). Furthermore, we also propose a large synthetic multiview multi-person HMR dataset that contains more people, more camera views, and scene variations (see Table 1 for reference) compared to existing datasets.

In summary, the contributions of the paper are:

- As far as we know, this is the first study on the multiview multi-person HMR task under large scenes with severe occlusions. No existing research has focused on the issue in the HMR area. Besides, we propose a large MVMP-HMR dataset for studying the topic.
- We propose the MVMP-HMR model, which is the first multiview multi-person HMR model for reconstructing multiple persons with multiple views under large scenes. In addition, we propose two novel losses for better MVMP-HMR performance.
- Experiments demonstrate that existing methods cannot perform well under the new multiview multi-person HMR benchmark with severe occlusions, and the proposed MVMP-HMR method outperforms both existing single-view HMR state-of-the-arts (SOTAs) and 3D HPE with multi-view settings.

2 RELATED WORK

Single-person HMR. Human mesh recovery (HMR) predicts the human 3D meshes from images. The early HMR methods were based on optimization, and they were easily stuck at local minima (Hasler et al., 2010; Lin et al., 2023; Moon et al., 2022; Pavlakos et al., 2019). Instead of estimating the human meshes as in 3D reconstruction, (Kanazawa et al., 2018) proposed to predict SMPL parameters of the shape and 3D joint angles to represent human meshes from a cropped image. SMPLify-X (Pavlakos et al., 2019) followed SMPLify to estimate the 2D joints and optimize model parameters to fit them, and then improved over SMPLify with a new DNN trained on a larger dataset. In addition, many regression-based methods were proposed (Cai et al., 2024a; Choutas et al., 2020; Feng et al., 2021; Moon et al., 2022; Rong et al., 2021; Zhang et al., 2023; Zhou et al., 2021), which is focused on single-person estimation. Furthermore, many methods tried to utilize multi-crops to

Table 1: The statistics of the proposed MVMP-HMR dataset, Single-view HMR, and 3D HPE datasets. MVMP-HMR dataset contains more persons, more scenes with multiviews, and more complexities.

Task	Dataset	Area	SceneNum	Subjects	Occlusion	Views	Frames	GT Format
Single-view HMR	GTA-Human	-	-	1	Simple	1	1.4M	**SMPL**, J3D
	EHF	-	-	1	Simple	1	100	**SMPLX**, J3D
	AGORA	-	-	5∼15	Medium	1	17K	**SMPLX**, SMPL, Mask
	BEDLAM	-	-	1∼10	Medium	1	380K	**SMPLX**
3D HPE	Human3.6M	4mx3m	7	1	Simple	4	3.6M	**SMPL**, J3D, Depth
	3DPW	-	-	1∼2	Simple	1	51K	**SMPL**
	CMU Panoptic	5.49mx4.15m	1	3∼8	Medium	65	1.5M	J3D, Depth
MVMP-HMR	Ours	30mx30m	15	10∼30	Severe	50	63K	**SMPLX**, J3D, Mask, Depth

enhance the HMR performance (Choutas et al., 2020; Feng et al., 2021; Moon et al., 2022; Lin et al., 2023; Cai et al., 2023). *In summary, single-person HMR is limited to images with few persons, making it impractical for real-world scenarios with multiple people, larger scenes, and severe occlusion.*

Multi-person HMR. Compared to single-person HMR, multi-person HMR (Choi et al., 2022a; Goel et al., 2023; Kolotouros et al., 2019; Qiu et al., 2022; Zhang et al., 2021a) needs to predict the human meshes of multiple persons in the images. Multi-person HMR usually adopts a two-stage procedure: detect all humans in the image first (He et al., 2017; Liu et al., 2016; Redmon et al., 2016), and then perform HMR (Kim et al., 2023; Ma et al., 2023; Yoshiyasu, 2023; Zheng et al., 2023) for each detected person with crops. The two-stage process is not end-to-end and the occlusion in images may hurt the human detection accuracy, thus limiting the whole pipeline's performance. In contrast, single-stage methods have also been proposed (Sun et al., 2021; Qiu et al., 2023; Sun et al., 2022). Recent methods Multi-HMR (Baradel et al., 2024) and AiOS (Sun et al., 2024) adopted the DETR architecture for multi-person human mesh recovery. Multi-HMR (Baradel et al., 2024) detects 2D people locations using features of a ViT backbone and predicts their whole-body pose, shape, and 3D location using a cross-attention module. AiOS (Sun et al., 2024) performs human localization and SMPL-X estimation in a progressive manner, which consists of body localization, body refinement, and a whole-body refinement stage to regress SMPL-X parameters. Even though existing multi-person HMR methods can accurately estimate human meshes for several persons in single images, they are only evaluated on small scenes containing a small number of persons, eg, < 15. It is not clear whether they can be applied to scenes with larger sizes and severe occlusions. Thus, we propose MVMP-HMR, which fuses multiple camera views to deal with severe occlusions. *As far as we know, this is the first study for multi-person HMR with multiviews, and we also propose a large synthetic MVMP-HMR dataset, which shall advance the HMR task to more complicated conditions.*

Single-view HMR and 3D HPE Datasets. While numerous datasets have been proposed for Human Mesh Recovery (HMR) and other 3D human tasks (*eg*., 3D Human Pose Estimation (HPE)), they have distinct human number, area size, and environmental complexity limitations compared with our dataset, as shown in Table 1. *Single-view HMR Datasets* like GTA-Human (Cai et al., 2024b), AGORA (Patel et al., 2021), and BEDLAM (Black et al., 2023) all employ synthetic data generation through game engines, and EHF (Pavlakos et al., 2019) is collected in the laboratory. Though providing SMPL-family parametric labels, they fundamentally suffer from depth ambiguity in monocular capture and lack real-world scene complexity. The number of people appearing in their scene is quite small, mostly just one person or at most 15 people in the scene, which is not practical in the real outdoors. Besides, since their scenes are quite simple with no other obstacles in the environment, the occlusion levels of the scenes are quite low. Therefore, existing HMR datasets are mainly based on single-view images, which are not applicable to more complicated scenes with large sizes and severe occlusions. *Compared to Single-view HMR datasets, our MVMP-HMR dataset provides a greater variety of views and a larger number of people. So, MVMP-HMR is more applicable in severe occlusion scenes.*

3D HPE Datasets include Human3.6M (Ionescu et al., 2013), 3DPW (Von Marcard et al., 2018), and CMU Panoptic (Joo et al., 2015a). While they capture real-world data through camera arrays and mocap systems, they still have three key limitations: 1) Limited human count: Typically each scene contains $\leq$10 subjects, failing to represent crowded real-world environments; 2) Constrained scene sizes and type: They are all captured in studio environments or small indoor spaces (leq50 m^2), lacking large-scale outdoor variations, their background type is limited in the indoor scene, and

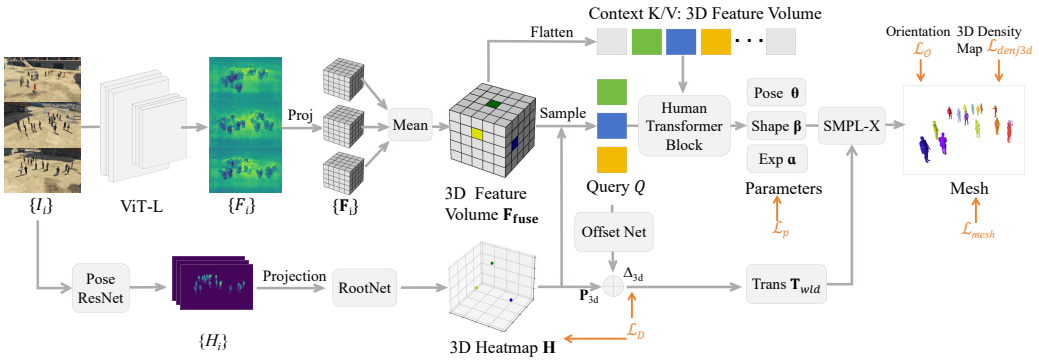

Figure 2: The pipeline of our proposed MVMP-HMR method, which consists of 3 main steps: Single-view Feature Extraction, Multi-view Feature Projection and Fusion, and 3D Decoding. We first extract single-view features with a ViT backbone, and then the single-view features are projected to the 3D space and averaged to obtain the 3D feature volume of the whole scene. Finally, with the joints outputted from a 3D pose estimation branch (at the bottom), we extract human queries for each person and feed them into a human transformer block (HTB) for 3D decoding and SMPL-X parameters prediction. In addition to losses previously used in single-view HMR SOTAs, we also put forward two novel losses, *eg.*, orientation loss $\mathcal{L}_{\mathcal{O}}$ and 3D joint density loss $\mathcal{L}_{denj3d}$, for better orientation and pose accuracy in meshes.

they cannot cover the light or time change outdoors; 3) Simplistic occlusion patterns: Due to the limited number of people, they primarily contain light inter-person occlusion. *3D HPE datasets have a fixed environment setting, while our MVMP-HMR dataset can simulate changes in lighting and provide more expansive scenes. MVMP-HMR also offers more extensive annotation than 3D HPE Datasets. These strengths make our dataset more representative of real-world scenarios and better suited for practical applications.*

3 Multiview Multi-Person HMR (MVMP-HMR)

We now introduce our multiview multi-person whole-body human mesh recovery task. Given multiview input RGB images $\mathbf{I} = \{I_1, I_2, \ldots, I_C\}$ (C is the view number), our model (denoted as **f**), directly predicts a group of N centered whole body SMPL-X parameters such as pose $\theta \in \mathbb{R}^{N\times 53\times 3}$, shape $\beta \in \mathbb{R}^{N\times 1\times 10}$, and expression $\alpha \in \mathbb{R}^{N\times 1\times 10}$, along with their associated 3D spatial translation $\mathbf{T}_{wld} \in \mathbb{R}^{N\times 1\times 3}$ in the world coordinate system. It outputs expressive 3D human meshes $\mathbf{M} = \textbf{SMPL-X}(\theta, \beta, \alpha, \mathbf{T}_{wld}) \in \mathbb{R}^{N\times 10475\times 3}$:

Compared to single-view human mesh recovery (Single-view HMR), MVMP-HMR task obtains human meshes with absolute locations in 3D world coordinates, rather than relative positions in the camera-view coordinates, because single-view prediction has depth, orientation, pose, and occlusion ambiguities. Thus, MVMP-HMR utilizes multiple views for better multi-view fusion and multi-person mesh recovery to deal with these ambiguities and severe occlusions in practical applications. We require the multiple cameras to be calibrated and synchronized in the setting. As in Figure 2, the proposed MVMP-HMR model consists of three modules: Single-view Feature Extraction, Multi-view Feature Projection and Fusion, and 3D Decoding, whose details are as below.

3.1 Single-view Feature Extraction

Our MVMP-HMR framework employs the Vision Transformer-Large (ViT-L) (Dosovitskiy et al., 2021) architecture as the backbone single-view feature extractor: $F_i = \textbf{ViT-L}(I_i)_{i\in\{1,\ldots,C\}}$, where i denotes the view id, F_i denotes the feature map of view I_i, and C is the number of views. To validate backbone selection, we conduct comprehensive experiments comparing various transformer-based architectures, with detailed ablation studies presented in the Appendix A. The ViT-L model demonstrates superior performance in capturing global contextual features critical for multi-view fusion. Thus, we use ViT-L as the feature extractor.

In parallel with the ViT-L backbone, we use an HRNet (Sun et al., 2019) for 2D pose heatmap predictions H_i.After the single-view feature extraction, we obtain feature maps $\{F_i\}$ and heatmaps $\{H_i\}$ of all views. They are forwarded to the next step for fusion.

3.2 Multi-view Feature Projection and Fusion

The extracted single-view features are projected to a constructed 3D volume for multiview feature fusion. The constructed 3D volume size is $300 \times 300 \times 20$, each voxel dimension representing 100mm in the physical 3D world. So the volume's spatial dimensions are $30\text{m} \times 30\text{m} \times 2\text{m}$ in the real world. In the feature projection, we employ perspective geometries to map each 3D voxel coordinate $\mathbf{p}_w = (x, y, z)$ to 2D image coordinates of multiple views: $\mathbf{p}_c^{(i)} = \mathbf{K}^{(i)}[\mathbf{R}^{(i)}|\mathbf{t}^{(i)}]\mathbf{p}_w$, where intrinsic $\mathbf{K}$ and extrinsic $[\mathbf{R} \mid \mathbf{t}]$ matrices are provided in the MVMP-HMR dataset, and i denotes the camera view index. We project each view's feature map F_i into a 3D volume through this perspective-aware coordinate projection, and each view's 3D feature volume is denoted as $\mathbf{F}_i$. Then, we fuse the projected multi-view feature volumes via a mean operation, and the fusion result is denoted as $\mathbf{F}_{\mathbf{fuse}}$.

2D heatmaps H_i are projected into a 3D volume, then fed into a modified RootNet (Tu et al., 2020) to generate 3D probability heatmaps $\mathbf{H}$ (encoding pelvis joint likelihoods in world coordinates). Fusion of these heatmaps yields the coarse 3D grid location $\mathbf{P}_{\mathbf{3d}}$ of the primary (pelvis) joint.

3.3 3D Decoding

The fused 3D feature volume $\mathbf{F}_{\mathbf{fuse}}$ is decoded with a Human Transformer Block (HTB) to regress the SMPL-X parameters in the 3D world. For each detected human $n \in \{1, ..., N\}$ in the 3D heatmap $\mathbf{H}$, we use pelvis joints to sample human features q from $\mathbf{F}_{\mathbf{fuse}}$. Then we combine q with X to construct human queries (denoted as Q), and X denotes the mean SMPL-X model parameters. Besides, the 3D feature volume $\mathbf{F}_{\mathbf{fuse}}$ is flattened as one-dimensional vectors as Keys and Values. Then we input Queries, Keys, and Values into our HTB for SMPL-X parameter regression.

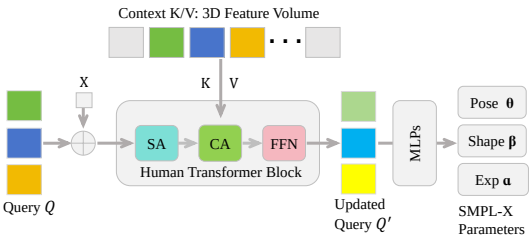

Figure 3: The details of the HTB: human queries are updated first via the self-attention layer (SA), the cross-attention layer (CA) integrated with flattened 3D features, and the FeedForward (FFN) layer, and then decoded via MLPs for SMPL-X parameter regression.

Figure 3 shows the details of the Human Transformer Block. The full flattened vectors are used as cross-attention keys K and values V. The human queries Q are updated with a stack of D HTB. Then, three MLPs are introduced to regress each human's SMPL-X parameters θ, β, and α with the updated human queries Q'.

Human queries Q are also fed into a 3D offset prediction net to estimate the offset Δ_{3d} of humans. Combining the primary joint location $\mathbf{P}_{\mathbf{3d}}$ in the 3D heapmap and Δ_{3d}, we can get the final location of the human's primary location, denoted as translation $\mathbf{T}_{wld} = \mathbf{P}_{\mathbf{3d}} + \Delta_{3d}$. Finally, we input the SMPL-X parameters and the translation $\mathbf{T}_{wld}$ to the SMPL-X layer (Pavlakos et al., 2019) for acquiring humans' mesh vertices and joints locations in world and camera view coordinates.

3.4 Training Loss

Overall, we adopt five losses to train the proposed MVMP-HMR model. The first three types of losses are similar as in the prior work (Baradel et al., 2024): the **detection loss** for localizing the human queries, the **SMPL-X parameter regression loss**, and the **mesh loss** for supervising 3D joints and vertices coordinate regression in human mesh format. Besides, since our task is in the 3D coordinates system, with orientation and pose ambiguities under the occluded scenes, we propose two novel losses to further enhance the model performance: the **orientation loss** for better orientation prediction instead of the direct SMPL-X parameters predictions, and the **3D joint density loss** supervising the predicted 3D joints from the human meshes in 3D density format instead of direct 3D joint coordinate regressing. The details of each loss are as follows.

Detection loss. With the help of the heatmap prediction branch **HRNet** (Sun et al., 2019), we can get the 3D heatmap $\mathbf{H}$ of the primary joint of each human in the scene. Then we construct a 3D volume to present the occupancy of people as $\hat{\mathbf{H}}$ with GT joints location. We also obtained the 3D offset Δ_{3d} in the grid to get a more refined coordinate. So we have the detection loss $\mathcal{L}_D$ as follows: $\mathcal{L}_D = ||\mathbf{H} - \hat{\mathbf{H}}||_2 + |\Delta_{3d} - \hat{\Delta}_{3d}|$. where $\hat{\mathbf{H}}$ and $\hat{\Delta}_{3d}$ are the ground truth 3D heatmap and location offset of the joints, respectively.

Parameter regression loss. All SMPL-X parameters predicted by the model are computed with L_1 regression losses. We integrate the body model parameters (pose θ, shape β, expression α) into loss function as follows: $\mathcal{L}_p = |\theta - \hat{\theta}| + |\beta - \hat{\beta}| + |\alpha - \hat{\alpha}|$, where $\hat{\theta}$, $\hat{\beta}$, and $\hat{\alpha}$ are the GT parameters.

Mesh loss. After predicting SMPL-X parameters, we can construct human meshes from a SMPL-X layer. Then we extract 3D joints J_{3D} and vertices V_{3D} from the human meshes and project these 3D points onto the 2D multi-image planes. The mesh loss supervises the 3D/2D vertices and joints:

$$\mathcal{L}_{3D} = |J_{3D} - \hat{J}_{3D}| + |V_{3D} - \hat{V}_{3D}|, \mathcal{L}_{2D} = |\pi_i(J_{3D}) - \pi_i(\hat{J}_{3D})| + |(\pi_i(V_{3D}) - \pi_i(\hat{V}_{3D})|, \quad (1)$$

where $\hat{J}_{3D}$ and $\hat{V}_{3D}$ are the ground truth 3D joints and vertices, π_i is the camera projection operator, and $\pi_i(\hat{J_{3D}})$ and $\pi_i(\hat{V_{3D}})$ refer to the ground truth 2D joints and vertices projected from the 3D ground truth. And the mesh loss $\mathcal{L}_{mesh}$ combines the two losses: $\mathcal{L}_{mesh} = \lambda_1 \mathcal{L}_{3D} + \frac{1}{C} \sum_{i=1}^{C} \mathcal{L}_{2D}$. Loss weight λ_1 adjusts the weight for the two loss terms and we use a fixed value $\lambda_1 = 100$ in all experiments. *In addition to these losses, we propose two novel losses:*

Orientation loss. The global orientation (a low-dimensional vector) in SMPL-X parameters cannot effectively supervise the orientation of the generated human mesh. Thus, we define the orientation of the human mesh through the joint points for better human mesh orientation supervision (see Figure 4). Specifically, a human's left hip $\hat{J_{lhip}}$ and right hip $\hat{J}_{rhip}$ can provide the direction of the x-axis, and a human's pelvis $\hat{J}_{pelvis}$ and spine $\hat{J}_{spine}$ can offer the direction of the y-axis. We use the cross product of the x-axis vector and the y-axis vector to obtain the ground truth orientation $\hat{\mathcal{O}}$ of the human body: $\hat{\mathcal{O}} = (\hat{J}_{lhip} - \hat{J}_{rhip}) \times (\hat{J}_{spine} - \hat{J}_{pelvis})$. In this way, we compute the orientation loss $\mathcal{L}_{\mathcal{O}}$ between the prediction joints $\mathcal{O}$ and ground-truth joints $\hat{\mathcal{O}}$ as: $\mathcal{L}_{\mathcal{O}} = |\mathcal{O} - \hat{\mathcal{O}}|$.

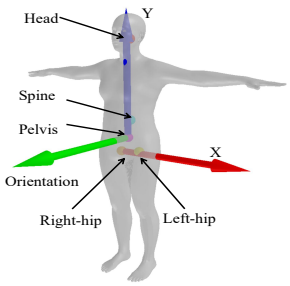

Figure 4: The orientation (green arrow) defined from human joints.

3D joint density loss. We use 3D Gaussian kernels to generate a density map of 3D joints from GT $\hat{J}_{3D}$ and prediction J_{3D}. Unlike the direct L_1 loss of 3D joint locations (as in mesh loss), we use mean square error loss (MSE) for the 3D density map regression:

$$\mathcal{L}_{denj3d} = \|\mathbf{Gau}(J_{3D}) - \mathbf{Gau}(\hat{J}_{3D})\|_2^2, \quad (2)$$

where **Gau** stands for the Gaussian smoothing step, which generates a 3D Gaussian probability map centered around the joint locations. The 3D joint density loss $\mathcal{L}_{denj3d}$ is conducted elemental-wisely in 3D space and provides stronger supervision for the pose of the human mesh, handling the pose ambiguities better in the MVMP HMR task under occlusions.

In total, the whole training loss is: $\mathcal{L} = \mathcal{L}_D + \lambda_2 \mathcal{L}_P + \mathcal{L}_{mesh} + \lambda_3 \mathcal{L}_{\mathcal{O}} + \lambda_4 \mathcal{L}_{denj3d}$. We set $\lambda_2 = 10$, $\lambda_3 = 5$ and $\lambda_4 = 1$ in our experiments.

4 Experiments and Results

4.1 Dataset

We perform the experiments on 3 datasets: MVMP-HMR collected by us, Panoptic (Joo et al., 2015b), and Human3.6M (Ionescu et al., 2013). The collection process of MVMP-HMR is as follows.

Dataset Generation. To study multiview multi-person human mesh recovery (HMR), we introduce MVMP-HMR, a large-scale dataset generated using the virtual game platform GTA-V. The dataset

Table 2: The result comparison on our MVMP-HMR, Human3.6M and Panoptic dataset. Rows 1-6 are single-view HMR SOTAs with multi-view fusion techniques, and Rows 7-8 are 3D pose estimation methods modified for SMPL-X regression.

Dataset	MVMR-HMR			Human3.6M			Panoptic		
Method	MPJPE ↓	PVE ↓	PA-PVE ↓	MPJPE ↓	PVE ↓	PA-PVE ↓	MPJPE ↓	PVE ↓	PA-PVE ↓
3DCrowdNet (Dist)	221.2	284.3	72.2	-	-	-	-	-	-
AiOS (Dist)	873.6	642.4	110.5	156.8	133.4	78.9	730.6	550.9	195.8
TokenHMR (Dist)	632.3	661.3	191.5	112.4	122.5	58.9	616.3	598.0	194.4
Multi-HMR (Dist)	841.0	651.4	71.0	98.5	97.3	46.3	568.7	453.4	195.1
Multi-HMR (Avg)	752.5	753.6	61.7	110.3	99.8	52.7	546.9	509.8	220.8
Multi-HMR (Fusion)	602.4	529.5	111.4	129.7	122.8	65.4	523.3	423.1	192.6
VoxelSMPLX (Only)	225.4	262.0	240.6	-	-	-	-	-	-
VoxelSMPLX (Joint)	288.6	427.4	317.1	-	-	-	-	-	-
MVMP-HMR (Ours)	**177.5**	**129.2**	**51.8**	**93.5**	**92.1**	**44.3**	**278.6**	**234.5**	**95.3**

features diverse everyday scenes (e.g., basketball courts, factories, streets) with varying numbers of people (10–30 per scene), complex occlusions, and up to 50 camera views per scene. Using GTA-VAPIs, we extract 98 3D body keypoints, depth maps, and semantic masks for each scene. In total, MVMP-HMR contains 15 complex scenes, making it the first large-scale multiview multi-person HMR dataset, designed to advance HMR research in challenging, real-world-like environments.

Dataset Annotation. Since GTA-V APIs do not provide 3D mesh labels, we adopt an HMR method (Baradel et al., 2024) for SMPL-X annotation in 3D world coordinates. To obtain accurate SMPL-X parameters, we first apply (Baradel et al., 2024) on all views of a frame to obtain SMPL-X labels in the camera coordinates of all people. *Then, for each person, we match the ground-truth 2D keypoints provided in GTA-V and the ones extracted from the predicted SMPL-X labels of all views. The SMPL-X label with the lowest matching error is assigned as the ground truth of the corresponding person*. In contrast to the single-view HMR task, the MVMP-HMR task estimates the human meshes in 3D world coordinates. Thus, we transform these 'predicted' ground-truth human meshes to world coordinates via a rotation and translation matrix.

From the single-view HMR prediction, we obtain global orientation R_{cam} and translation T_{cam} to decide the directions and locations of the human mesh in camera coordinates. We then compute R, T between 3D joint points shared in predicted SMPL-X mesh format (camera coordinates) and GTA-V (world coordinates). Then ground truth (GT) global orientation parameter R^{gt}_{wld} and translation parameter T^{gt}_{wld} are formulated as: $R^{gt}_{wld} = R \cdot R_{cam}$ and $T^{gt}_{wld} = T_{cam} + T$. The SMPLX annotation acquisition for the real dataset Panoptic (Joo et al., 2015b) is consistent with the above content. SMPLX label in Human3.6m are obtained from Choi et al. (2022b)

4.2 Experiment Settings

Implementation. In experiments, we divide the 15 scenes in the dataset according to the distribution of people numbers, and the ratio of the training/testing set is 2:1. We use VIT-L (Dosovitskiy et al., 2021) as our model feature extraction backbone. We pre-train the posenet (Sun et al., 2019) and rootnet (Tu et al., 2020) for 60 epochs on our dataset for detection. The input images are resized to 1288 x 1288 with zero paddings. We adopt Adam as the optimizer with 5e-5 learning rate. The training epoch is 50, and the training is conducted on 2 RTX6000 Ada GPUs, with a batch size of 1.

Comparison methods. We compare our MVMP-HMR method with multi-person HMR SOTAs with multiview settings and 3D HPE method for HMR tasks. Single-view HMR SOTAs Multi-HMR (Baradel et al., 2024), 3DCrowdNet (Choi et al., 2022b), AiOS (Sun et al., 2024), and TokenHMR (Dwivedi et al., 2024) first conduct predictions of each view, then use a multi-view matching algorithm to match the prediction results of each person under multiple views, and fuse the prediction results of each person in the scene under multiple views into the final result. The fusion strategy includes selecting the closest one as the prediction result based on the distance from the camera (denoted as 'Dist'), using an average strategy to fuse the results of each view prediction (denoted as 'Avg'), and using a sub-network to predict the weight value corresponding to each view prediction to fuse the final result (denoted as 'Fusion'). We also compare with a multi-view 3D pose estimation method VoxelPose (Tu et al., 2020). We sample human queries from the feature volume with the predicted joint locations of VoxelPose (Tu et al., 2020) and then estimate the SMPL-X parameters from the

Figure 5: The top row is the multiview input, and each subsequent row is the 3D predictions of the methods projected to view plane. Red boxes indicate that our method can better handle occlusions than comparison methods. Blue boxes indicate our method achieves better posture than comparisons.

human queries with regression MLPs. There are two variants: use the pretrained VoxelPose and only train the regression MLPs, denoted 'VoxelSMPLX (Only)'; or jointly train VoxelPose and MLPs, denoted as 'VoxelSMPLX (Joint)'.

Table 3: Loss term ablation study. The first row does not use any new loss, the second row only adds the orientation loss, the third row only adds the 3D joint density loss, and the last row adds both new losses (our method).

Loss	MPJPE ↓	PVE ↓	PA-PVE ↓
$\mathcal{L}_D+\lambda_2\mathcal{L}_P+\mathcal{L}_{mesh}$	217.1	161.7	120.4
$+5.0\mathcal{L}_{\mathcal{O}}$	187.9	144.8	89.0
$+1.0\mathcal{L}_{denj3d}$	180.7	132.4	50.2
+Both (Ours)	**177.5**	**129.2**	**51.8**

4.3 MVMP HMR Results

We comprehensively evaluate our MVMP-HMR model against state-of-the-art approaches on three benchmarks: MVMP-HMR (synthetic), Human3.6M, and CMU Panoptic Dataset in Table 2. The comparison includes six single-view HMR baselines enhanced with multi-view fusion techniques, such as 3DCrowdNet (Dist), AiOS (Dist), TokenHMR (Dist), Multi-HMR (Dist), Multi-HMR (Avg), and Multi-HMR (Fusion), and a 3D human pose estimation (3D HPE) method added with SMPL-X regression net, eg., VoxelSMPLX (Only) and VoxelSMPLX (Joint). According to Table 2, we conclude that the proposed MVMP-HMR model achieves much better results than all comparisons.

Table 4: Feature fusion method ablation study.

Fusion Method	MPJPE ↓	PVE ↓	PA-PVE ↓
Deformable	261.3	207.2	80.6
Max	245.2	193.5	74.8
Mean (Ours)	**177.5**	**129.2**	**51.8**

Table 5: Primary joint ablation study.

Primary Joint	MPJPE ↓	PVE ↓	PA-PVE ↓
Head	280.6	172.3	68.2
Spine	190.2	146.9	86.1
Pelvis (Ours)	**177.5**	**129.2**	**51.8**

The reason is that these comparison methods are primarily designed for single-view HMR or 3D HPE in simple scenes with only a few humans. The former cannot effectively fuse multiview clues to handle occlusions, while the latter cannot accurately estimate human meshes solely from 3D poses, lacking useful shape information. This demonstrates the advantages of the proposed MVMP-HMR model in handling severe occlusions and human orientation or pose ambiguities in complex scenes.

As **visualized** in Figure 5, our proposed method outperforms all comparison methods, in terms of predicting completeness (no person is missed) and pose accuracy. The *red boxes* indicate our method can handle occlusions well and estimate meshes accurately for occluded persons, while all comparisons neglect the occluded persons or produce wrong shapes. The *blue boxes* indicate our method achieves more natural and realistic human poses, with better limb positioning and alignment compared to comparison methods that produce unrealistic limb orientations and poses (such as flying pose in the first row, fourth column of 3DCrowdNet (Dist), hugging posture in the six row, third column of Multi-HMR (Fusion), or VoxelSMPLX).

4.4 Ablation Study

Loss term ablation study. We conduct ablation studies on two novel losses—orientation Loss $\mathcal{L}_{\mathcal{O}}$ and 3D joint density Loss $\mathcal{L}_{denj3d}$ —by incorporating them individually or together with three standard single-view HMR losses. As shown in Table 3, both new losses improve the performance of our MVMP-HMR model, and using both together achieves the best results, demonstrating their effectiveness in reducing orientation and pose ambiguities in multiview multi-person HMR. Notably, $\mathcal{L}_{denj3d}$ contributes more significantly, providing stronger 3D supervision and greater performance gains. **See detailed loss term weight ablations in Table 8 of the Appendix.**

Feature fusion method ablation study. We also perform ablation studies on the feature fusion method, using three different methods: Deformable attention, Max, and Mean. As in Table 4, the performance using the mean operation to fusion multi-view features achieves marginally superior performance than using deformable attention or max. The possible reason is that the mean method is simple and efficient, suitable for global information fusion, but max is suitable for highlighting key features, but is susceptible to noise interference. And the deformable attention has a high computational overhead. In our setting, the mean operation is better for our environment to aggregate multi-view features. Thus, in our experiments, we use the mean as the feature fusion method.

Primary joint selection ablation study. To determine the optimal primary joint for our model, we conducted an ablation study comparing three different primary joints: the pelvis, head, and spine. As in Table 5, the results show that the use of the pelvis for localisation produces marginally better performance. This can be attributed to the pelvis's stability across various viewpoints and its central location, which allows for more complete human body information to be captured in the model's queries. Consequently, we chose the pelvis as the primary joint for all subsequent experiments. **See model architecture and view number ablations in the Appendix**.

5 Conclusion

In this paper, we propose a novel multi-person whole-body human mesh recovery model from multiview images and a new large multiview HMR benchmark with more persons in large occluded scenes. As far as we know, this is the first study on multiview-multiperson-based (MVMP) HMR tasks and the first large MVMP-HMR benchmark in this area. Besides, two novel losses are put forward to further enhance the model's performance: the orientation loss and the 3D joint density loss, handling the orientation and pose ambiguities in the mesh predictions under the occluded scenes. The experiments validate that the MVMP-HMR model can deal with the occlusion issue better than existing single-view HMR SOTAs. The proposed model and benchmark shall extend the HMR task to more complicated scenes with wider application scenarios.

Ethics Statement

This work introduces a framework for multiview multi-person human mesh recovery (MVMP-HMR) using a synthetic dataset generated with the GTA-V engine and publicly available benchmarks such as Human3.6M (Ionescu et al., 2013) and CMU Panoptic (Joo et al., 2015b), all of which contain no personally identifiable information. SMPL-X annotations are derived automatically using existing HMR models, reducing the need for manual labeling and associated privacy concerns. Our research advances human mesh recovery with potential benefits in motion analysis, human-computer interaction, and safety-critical applications. While we are not aware of negative societal impacts specific to our method, we acknowledge broader ethical considerations related to surveillance, fairness, and potential misuse, and emphasize responsible and transparent deployment.

Reproducibility Statement

We provide detailed descriptions of our MVMP-HMR architecture, including the ViT-L (Dosovitskiy et al., 2021) backbone, multi-view feature fusion, and Human Transformer Block, along with the proposed orientation and 3D joint density losses. Implementation details such as training configuration, hyperparameters, and dataset splits are reported in Section 4.2. Experiments are conducted on 2 NVIDIA RTX 6000 Ada GPUs, and we will release the MVMP-HMR dataset, source code, pre-trained models, and training logs upon acceptance, ensuring reproducibility and facilitating future research.

References

Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *European Conference on Computer Vision*, pp. 202–218. Springer, 2024.

Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8726–8737, 2023.

Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36, 2024a.

Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, Chen Change Loy, and Ziwei Liu. Playing for 3d human recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024b.

Zizhuo Cai, Wenhao Yin, Ailin Zeng, Chunhe Wei, Qi Sun, Yezhi Wang, Hui En Pang, Hongyu Mei, Mingmin Zhang, Liangjian Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

H. Choi, G. Moon, J. Park, and K.M. Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *CVPR*, 2022a.

Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1475–1484, 2022b.

Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pp. 20–40. Springer, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1323–1333, 2024.

Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *2021 International Conference on 3D Vision (3DV)*, pp. 792–804. IEEE, 2021.

S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *ICCV*, 2023.

Nils Hasler, Hanno Ackermann, Bodo Rosenhahn, Thorsten Thormählen, and Hans-Peter Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1823–1830. IEEE, 2010.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

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.

Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE international conference on computer vision*, pp. 3334–3342, 2015a.

Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3334–3342, 2015b.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7122–7131, 2018.

J. Kim, M.G. Gwon, H. Park, H. Kwon, G.M. Um, and W. Kim. Sampling is matter: Point-guided 3d human mesh reconstruction. In *CVPR*, 2023.

N. Kolotouros, G. Pavlakos, M.J. Black, and K. Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019.

Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21159–21168, 2023.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, 2016.

X. Ma, J. Su, C. Wang, W. Zhu, and Y. Wang. 3d human mesh estimation from virtual markers. In *CVPR*, 2023.

Yuto Matsubara and Ko Nishino. Heatformer: A neural optimizer for multiview human mesh recovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6415–6424, 2025.

Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2308–2317, 2022.

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*, pp. 13468–13478, 2021.

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*, pp. 10975–10985, 2019.

Z. Qiu, Q. Yang, J. Wang, and D. Fu. Dynamic graph reasoning for multi-person 3d pose estimation. In *ACMMM*, 2022.

Z. Qiu, Q. Yang, J. Wang, H. Feng, J. Han, E. Ding, C. Xu, D. Fu, and J. Wang. Psvt: End-to-end multi-person 3d pose and shape estimation with progressive video transformers. In *CVPR*, 2023.

Joseph Redmon, Santosh Divakaran, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1749–1759, 2021.

Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5693–5703, 2019.

Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi-Sing Leung, Ziwei Liu, Lei Yang, et al. Aios: All-in-one-stage expressive human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1834–1843, 2024.

Y. Sun, Q. Bao, W. Liu, Y. Fu, M.J. Black, and T. Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021.

Y. Sun, W. Liu, Q. Bao, Y. Fu, T. Mei, and M.J. Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, 2022.

Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 197–212. Springer, 2020.

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)*, pp. 601–617, 2018.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025.

Tao Wang, Jianfeng Zhang, Yujun Cai, Shuicheng Yan, and Jiashi Feng. Direct multi-view multi-person 3d human pose estimation. *Advances in Neural Information Processing Systems*, 2021.

Hang Ye, Wentao Zhu, Chunyu Wang, Rujie Wu, and Yizhou Wang. Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In *European Conference on Computer Vision (ECCV)*, 2022.

Y. Yoshiyasu. Deformable mesh transformer for 3d human mesh recovery. In *CVPR*, 2023.

Zhixuan Yu, Linguang Zhang, Yuanlu Xu, Chengcheng Tang, Luan Tran, Cem Keskin, and Hyun Soo Park. Multiview human body reconstruction from uncalibrated cameras. *Advances in Neural Information Processing Systems*, 35:7879–7891, 2022.

H. Zhang, 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. In *ICCV*, 2021a.

Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12287–12303, 2023.

Qi Zhang, Wei Lin, and Antoni B Chan. Cross-view cross-scene multi-view crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 557–567, 2021b.

C. Zheng, X. Liu, G.J. Qi, and C. Chen. Potter: Pooling attention transformer for efficient human mesh recovery. In *CVPR*, 2023.

Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular real-time full body capture with inter-part correlations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4811–4822, 2021.

A APPENDIX

A.1 METRIC DETAILS.

We evaluate the HMR predictions with metrics MPJPE, PVE, and PA-PVE, but **in 3D space**, not in camera view as previous single-view HMR tasks.

- *MPJPE*: Mean Per Joint Position Error measures the average Euclidean distance between predicted 3D joints and ground truth 3D joints.
- *PVE*: Mean Per-vertex Error is defined as the average point-to-point Euclidean distance between predicted mesh vertices and ground truth mesh vertices. It is proposed to calculate in the world space.
- *PA-PVE*: Procrustes-aligned PVE is calculated according to PVE after executing Procrustes Analysis to align predicted mesh vertices with ground truth mesh vertices.

MPJPE and PVE are the main metrics in our task.

A.2 MAIN FEATURES OF MVMP-HMR VS SOTA HMR AND HPE METHODS.

As shown in Table 6, we have compared over 10 human mesh recovery and human pose estimation methods. It is easy to see that our method is the only one that focuses on the multiview multi-person human mesh recovery task.

Table 6: Comparison of existing HMR and HPE Methods. None of them meets our setting without revision.

Method	**Multi-view**	**Multi-person**	**3D Pose**	**Mesh**
HMR(Kanazawa et al., 2018)	×	×	✓	✓
PyMAF-X(Zhang et al., 2021a)	×	×	✓	✓
OSX(Lin et al., 2023)	×	×	✓	✓
SMPLer-X(Cai et al., 2023)	×	×	✓	✓
3D CrowdNet(Choi et al., 2022b)	×	✓	✓	✓
AiOS(Sun et al., 2024)	×	✓	✓	✓
TokenHMR(Dwivedi et al., 2024)	×	✓	✓	✓
Multi-HMR(Baradel et al., 2024)	×	✓	✓	✓
U-HMR(Yu et al., 2022)	✓	×	✓	✓
HeatFormer(Matsubara & Nishino, 2025)	✓	×	✓	✓
VoxelPose(Tu et al., 2020)	✓	✓	✓	×
Faster VoxelPose(Ye et al., 2022)	✓	✓	✓	×
MVP(Wang et al., 2021)	✓	✓	✓	×
MVMP-HMR (Ours)	✓	✓	✓	✓

Note: ✓ indicates supported, × indicates not supported.

A.3 DATASET

The GTA-V game engine demonstrates exceptional authenticity and has been widely adopted for dataset generation across various research fields, including GTA-Human (Cai et al., 2024b) and the multi-view counting dataset CVCS(Zhang et al., 2021b), offering highly realistic scenes, dynamic weather systems, comprehensive lighting variations, and diverse human activities such as walking, phone usage, drinking, smoking, listening to music, and social interactions. Our dataset shows a strong bias toward clear/sunny conditions (78.12%) with overcast coverage (12.78%) and adverse weather (9.09%), while temporal distribution exhibits pronounced daytime bias (79.73% between 6:00-18:00) with activity peaks during commuting hours and sparse nighttime coverage (20.27%). Compared to traditional 3D HPE datasets that are primarily collected in controlled laboratory settings, our GTA-V-generated dataset focuses on outdoor practical application scenarios with broader scene diversity and enhanced ecological validity, better representing the complexity and variability encountered in real-world conditions. More detailed dataset analysis to be added to the paper.

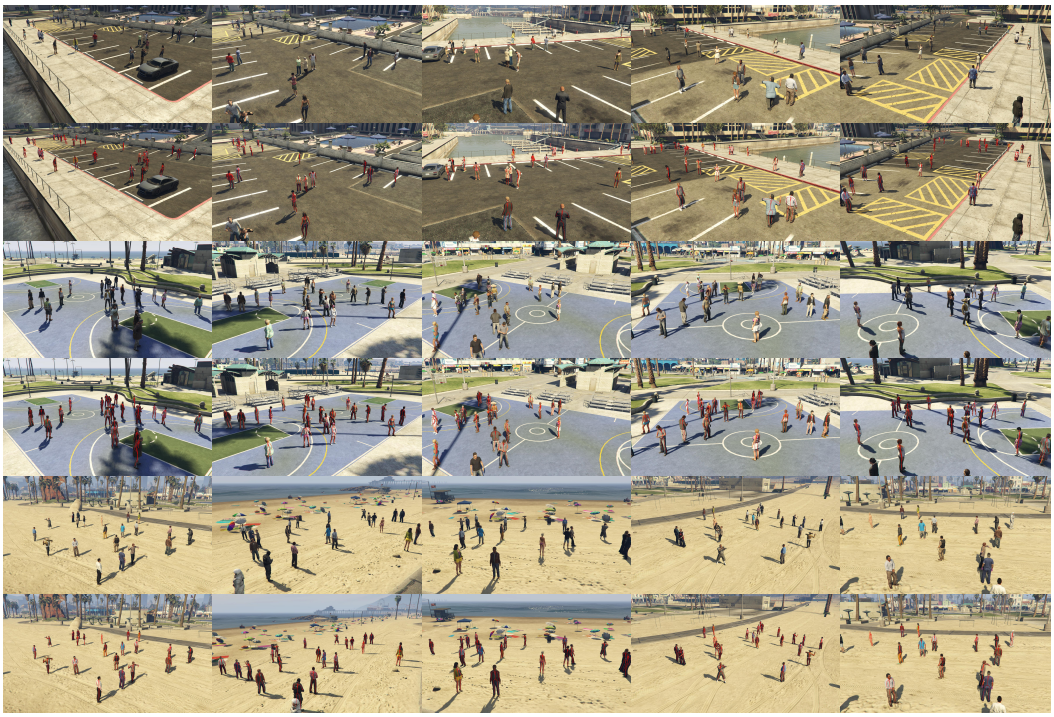

Figure 6: The visualization examples of the other scenes in the dataset. Red joints mean the keypoints of humans. The red line means the skeleton of people.

A.4 More visualizations of our MVMP-HMR dataset

We have introduced the dataset MVMP-HMR in the main text. Now we will show some other scenes in our dataset with their cooperation 3D joints, which are also key points for our dataset annotation. Figure 6 shows three scenes in our dataset. We can see the details of the 2D key-points location with red color. In our setting, the one who can't be seen completely at this view, their keypoint location will be dropped. These 2D keypoints are all projected from 3D keypoints. We can also provide precise keypoint locations for the multiview pose estimation task.

A.5 Model parameters and inference time

In addition to the results displayed in the dataset compared with other methods, we also made comparisons regarding model parameters and inference speed in Table 7. Our model parameters only count the model parameters during testing. Our inference time calculation is to run the model for 100 sample inputs and then test the entire test set for an average test time. From our model framework, it can be seen that the 3D voxel features constructed from multi-view feature projections and fusion, as well as the subsequent network processing, are very resource-intensive. However, our model's parameters and inference speed achieve a moderate result compared to single-view HMR and multi-view HPE methods. Although the HPE method has a simpler network architecture, resulting in lower estimated model parameters and inference speed than ours, the HPE method can't achieve good results on our MVMP-HMR dataset. Single-view HMR does not involve the fusion of multi-view features, so its model parameter count is smaller than ours. Additionally, the efficiency of detecting directly on 3D voxel features is higher than that of multi-view matching, leading to shorter inference times for our method.

A.6 Loss term weight ablation study.

We conduct the loss term weight ablations for the proposed orientation loss ($\mathcal{L}_{\mathcal{O}}$) and 3D joint density loss ($\mathcal{L}_{denj3d}$) in Table 8. The first row uses the loss usually used in prior work (Baradel et al., 2024). Row 2-5 add the proposed orientation loss $\mathcal{L}_{\mathcal{O}}$ with different λ_3 weights, and the performance all improved compared to without it, demonstrating the effectiveness of the $\mathcal{L}_{\mathcal{O}}$ loss. $\lambda_3 = 5.0$ achieves

Table 7: The model parameters and inference time compared to HMR and HPE SOTAs.

Method	Model Parameters (MB) ↓	Inference Time (s) ↓
3DCrowdNet (Dist) (Choi et al., 2022b)	931.92	1.12
AiOS (Dist) (Sun et al., 2024)	1122.28	0.97
TokenHMR (Dist) (Dwivedi et al., 2024)	2598.57	2.44
Multi-HMR (Dist) (Baradel et al., 2024)	1210.17	2.33
Multi-HMR (Avg) (Baradel et al., 2024)	1210.17	2.33
Multi-HMR (Fusion) (Baradel et al., 2024)	1331.19	2.53
VoxelSMPLX (Only) (Tu et al., 2020)	404.45	1.00
VoxelSMPLX (Joint) (Tu et al., 2020)	404.45	1.00
MVMP-HMR (Ours)	1380.28	1.59

Table 8: Loss term weight ablation study. The first row does not use any new loss. Rows 2-5 add the orientation loss, and Rows 6-11 add both the orientation and 3D joint density loss.

Loss	MPJPE ↓	PVE ↓	PA-PVE ↓
$\mathcal{L}_D + \lambda_2 \mathcal{L}_P + \mathcal{L}_{mesh}$	217.1	161.7	120.4
$+2.0 * \mathcal{L}_{\mathcal{O}}$	201.6	151.9	99.4
$+\mathbf{5.0} * \mathcal{L}_{\mathcal{O}}$	187.9	144.8	89.0
$+10.0 * \mathcal{L}_{\mathcal{O}}$	195.1	149.2	80.7
$+100.0 * \mathcal{L}_{\mathcal{O}}$	195.0	150.0	71.8
$+5.0 * \mathcal{L}_{\mathcal{O}} + 0.1 * \mathcal{L}_{denj3d}$	190.2	144.5	89.4
$+5.0 * \mathcal{L}_{\mathcal{O}} + 0.2 * \mathcal{L}_{denj3d}$	190.7	149.7	83.3
$+5.0 * \mathcal{L}_{\mathcal{O}} + 0.5 * \mathcal{L}_{denj3d}$	187.6	147.4	87.9
$+\mathbf{5.0} * \mathcal{L}_{\mathcal{O}} + \mathbf{1.0} * \mathcal{L}_{denj3d}$ (**Ours**)	**177.5**	**129.2**	**51.8**
$+5.0 * \mathcal{L}_{\mathcal{O}} + 2.0 * \mathcal{L}_{denj3d}$	293.3	149.5	69.6
$+5.0 * \mathcal{L}_{\mathcal{O}} + 5.0 * \mathcal{L}_{denj3d}$	368.8	160.0	69.4

the best results, and we use it as the loss weight of $\mathcal{L}_{\mathcal{O}}$ in the experiments. Row 6-11 further add the proposed 3D joint density loss $\mathcal{L}_{denj3d}$ in the model training. $\lambda_3 = 5.0, \lambda_4 = 1.0$ achieves the best results. When λ_4 is too large, the 3D joint density loss may decrease the human mesh prediction performance because $\mathcal{L}_{denj3d}$ might be too strong.

Table 9: The backbone ablation study and using ViT-L is the best

Backbone	MPJPE↓	PVE ↓	PA-PVE ↓
ViT-S (Dosovitskiy et al., 2021)	201.6	157.8	64.9
ViT-B (Dosovitskiy et al., 2021)	185.7	141.8	61.6
ViT-L (Dosovitskiy et al., 2021)	**177.5**	**129.2**	**51.8**

A.7 Feature extraction backbone model

We also perform ablation studies on the feature extraction backbone models, using three different feature extraction backbone models: ViT-S, ViT-B, and ViT-L (Dosovitskiy et al., 2021), differing in model sizes: small, base, and large. As in Table 9, the result of using ViT-L as the backbone model is the best, which has more model parameters with stronger feature extraction ability. Therefore, we use ViT-L as the feature backbone model in our MVMP-HMR model.

A.8 Testing view number ablation study

Finally, we perform ablation studies on the input camera view number in the testing stage. The model is trained with 5 camera views and tested with different camera views, ranging from 3-9 camera views, shown in Table 10. We observe that as the testing camera view number increases, the model's performance also improves. The reason is that with more camera views, more clues are provided, and the proposed Multiview-HMR model can effectively fuse multiview information to handle the occlusions in the scene. The model performance change is not quite large when the camera view number decreases, also indicating our model's robustness to different view numbers.

Table 10: Testing camera view number ablation study: the model is trained on 5 views and tested with 3-9 views.

ViewNum	MPJPE ↓	PVE ↓	PA-PVE ↓
3	193.6	137.6	50.9
5	177.5	129.2	51.8
7	171.0	125.2	48.2
9	**168.1**	**122.0**	**47.9**

A.9 LIMITATIONS

In our experimental setting, we require the input to be multiple cameras that have been calibrated to obtain the internal and external parameters of the camera. Although this is difficult to obtain in the real world, many existing excellent multi-view matching algorithms (such as (Schönberger et al., 2016)) or VGGT (Wang et al., 2025) can perform camera calibration through multiple perspectives, which provides great help for the future application of our method. In the future, we can consider how to use multi-view without camera parameter calibration to perform multiview multi-person human mesh recovery.