Supplementary Materials: ANTHROPOS-V: benchmarking the novel task of Crowd Volume Estimation

We supplement the main paper by outlining further notes on the SMPL fitting process and an additional experiment on estimating volumes of single body parts (Sec. 1, 2). We complement Sec. 3.2 of the main paper with additional remarks on the task's evaluation metrics (Sec. 3, 4). In addition, we show that ANTHROPOS-V can also serve as a benchmark for the tasks of *Crowd Counting* and *Human Mesh Recovery* (HMR) (Sec. 5). Then, we illustrate more details on the implementation of baselines (Sec. 6), and we provide additional qualitative results, encompassing both success and failure cases on real and synthetic images (Sec. 7, 8). Additionally, we include some sample images from ANTHROPOS-V (Sec. 9). Finally, we present a cross-dataset evaluation, other remarks on CVE vs. Crowd Counting, and a tentative approach to leverage temporal information in CVE (Secs. 10, 11, 12).

1. Further notes on the SMPL fitting process

The process of fitting SMPL meshes to characters, particularly in complex environments such as the Grand Theft Auto V (GTA-V) game, involves a complex combination of techniques from 3D modeling, computer vision, and machine learning.

We begin by collecting all the pre-existent meshes in the GTA-V game. The characters are identified by a name and a list of eleven variations that, in turn, express the contingent appearance of the character. It is worth noting that characters with the same name and different appearances do not necessarily share the same volume. Hence, we fit an SMPL mesh to all characters' variations appearing in each scene. Initially, our fitting method retrieves characters' data, including their 3D models and texture information. The 3D models are then converted into the widely-used OBJ format (see Fig. 2, the first image of each sequence) accompanied by MTL files, which are required for defining the materials and textures of the model. As in [12], our objective is to achieve a tight fit that closely conforms to exposed bare-skin body parts such as the head or uncovered arms. Simultaneously, we seek a more relaxed fit in clothed body regions to diminish the impact of the added thickness introduced by clothing on the overall body volume. To perform this fitting process, we need both a 3D pose prior and knowledge of which vertices in the GTA-V mesh represent skin or clothing. Thus, we initiate the process by generating 10 visual renders of the GTA-V 3D characters. This is achieved by moving the camera around the textured 3D mesh of the characters, as depicted in the first line of Fig. 1.

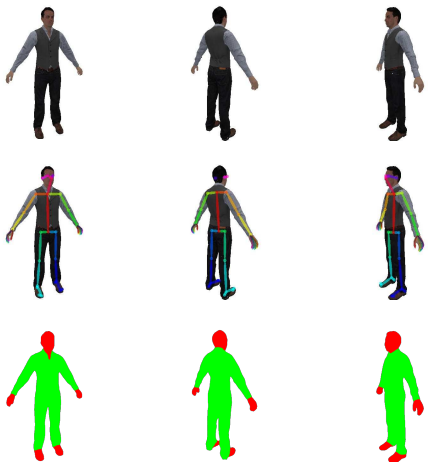


Figure 1. Renders from the SMPL mesh fitting process on a single character. The first line represents the renders, the second shows the estimated 2D pose for each render and the third is the Graphonomy output of skin segmentation (marked in red), as opposed to the dressed body segmentation (marked in green). Shoes are forced to be "skin" points to improve the fitting.

Then, the pose estimation process exploits [2] to predict the character's 2D pose in each rendered image, as depicted in the second line of Fig. 1. This 2D pose data is lifted into a three-dimensional space, giving a complete spatial representation of the character's posture. Next, the process of dividing a character's mesh into skin and clothes vertices leverages [5] to segment each of the 10 renderings (see Fig. 1, third line). The resulting segmentation is reprojected onto the mesh to label each vertex.

Before fitting the SMPL mesh, the character's gender is determined, which ensures the accuracy of the SMPL model, as these models are gender-specific. The SMPL fitting involves aligning a standard human body model to the character's 3D pose and shape. This step requires meticu-

Figure 2. Qualitative outcomes of the SMPL fitting process. Each image features: the original GTA-V character (first mesh), the output of the SMPL fitting process (second mesh), and an overlap of the GTA-V character with the SMPL result, in a front-facing view (third mesh), and in a backward-facing view (fourth mesh).

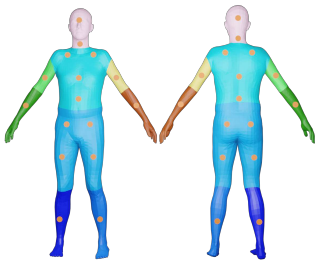


Figure 3. A SMPL mesh from ANTHROPOS-V. Different body parts are highlighted in different colors. Orange dots represent the keypoints associated with each body part.

lous adjustments to ensure that the SMPL mesh accurately follows the contours and posture of the character. Indeed, following [12], we employ two different loss functions to constrain the SMPL within the GTA-V mesh. The first loss is applied to the retrieved skin vertices, where we impose a severe fitting. The second applies to the clothes vertices, where we aim to have a looser fit so that these vertices would not penetrate the original mesh while remaining sufficiently close to it. Once the fitting process is over, the volume of the fitted SMPL mesh is computed using Blender [3] Python API, which calculates the volume within a mesh. A qualitative assessment of our fitting is present in Fig. 2. It is worth noting that our SMPL meshes typically lie beneath the attire of the GTA-V characters, with minimal penetration occurring primarily at skin vertices. This penetration is expected since we want a tighter fit in these specific areas.

Furthermore, to assign volume labels to individual body segments, we partition SMPL meshes into nine 3D parts, as illustrated in Fig. 3. The segmentation process relies on the body segmentation mappings presented in [10]. These mappings provide the indices of vertices corresponding to each body part, enabling the identification of boundary vertices situated between adjacent body parts. We split the meshes into disjointed body parts along these identified boundaries. Since boundary vertices often do not lie on a common plane, we identify the plane that traverses the maximum number of them while reporting the minimum distance from the non-traversed boundary vertices. Finally, we employ these planes to split the meshes into distinct body parts and compute their volumes.

2. Further Analysis of Crowd Volume Estimation Models

Metric	**Body Part**					
	Head	Arms	Forearms	Torso	Thighs	Calves
MAE	19.161	29.230	55.504	398.73	90.072	280.22
PP-MAE	1.1036	1.6920	3.6455	21.504	6.7060	20.707

Table 1. Volume error for each part of the body. The results are reported in dm^3.

To further investigate the performance of STEERER-V on Crowd Volume Estimation, we conducted additional experiments to assess its ability to localize volume within images. Specifically, we divided the images into random-scale patches and evaluated whether STEERER-V could accurately allocate the correct volume to each patch. The results (MAE: 130.2, PP-MAE: 5.8) are consistent with those from our main experiments, demonstrating that the model effectively distributes volume across the correct individuals.

Additionally, we assess STEERER-V capability to estimate the volume of single body parts. Specifically, we train our proposed model to estimate the volume of the single body parts' split presented in Sec. 4.3 of the main paper. Results of this experiment are reported in Table 1. While the error on the estimated volume of the head and arms is

Model	RMSE
CLIFF [8]	862.4
BEDLAM-CLIFF [1]	827.1
ReFit [17]	708.3
Oracular CLIFF [8]	473.9
Oracular BEDLAM-CLIFF [1]	459.5
Oracular ReFit [17]	412.7
$C_{B+}(I) \times \bar{V}_{\mathcal{D}}$	638.29
Bayesian+ [11]	904.23
P2P [15]	743.81
MAN [9]	915.64
STEERER [6]	643.10
Oracular $C(I) \times \bar{V}_{\mathcal{D}}$	254.91
STEERER-V [6]	**269.39**

Table 2. Results on ANTHROPOS-V, reported in dm^3. Methods are divided into HD+HMR, Crowd Counting, and our proposed approach. Gray-out lines rely on some oracular information and shouldn't be directly compared with the other results.

low, other parts like the torso and thighs expose a greater error due to a superior volume occupancy and loose-fitting clothes, rendering the correct estimation more challenging. Other body parts like calves and forearms have a high probability of being partially occluded or self-occluded, leading to a higher error compared to body parts with nearly the same volume coverage.

3. Further notes on the metrics

In Sec. 3.2 of the main manuscript, we introduced the minimal set of metrics for the Crowd Volume Estimation (CVE) task, particularly *Mean Absolute Error* (MAE) and *Per-Person Mean Absolute Error* (PP-MAE). We are aware that some literature on the task of *Crowd Counting* also reports the *Root Mean Squared Error* (RMSE). We argue that given a set of images $\{I_k\}$, RMSE is redundant for CVE, as it is proportional to MAE. We show this in Eq. 1, where $\{V_k\}$ is the total volume associated with each image, $\{\hat{V}_k\}$ the estimated one, and $\mathrm{AE}(k)$ is the absolute error of the k-th image.

$$\begin{aligned}\mathrm{RMSE}(\{I_k\}) &= \sqrt{\frac{1}{K}\sum_{k=1}^{K}(V_k - \hat{V}_k)^2} \\ &= \frac{1}{\sqrt{K}}\sqrt{\sum_{k=1}^{K}|V_k - \hat{V}_k|^2} \\ &= \frac{1}{\sqrt{K}}\sqrt{\sum_{k=1}^{K}[\mathrm{AE}(k)]^2} \\ &\propto \frac{1}{K}\sum_{k=1}^{K}\mathrm{AE}(k) = \mathrm{MAE}(\{I_k\})\end{aligned} \tag{1}$$

Nonetheless, we extend Table 1 of the main paper with Table 2, showing the RMSE for the proposed models.

4. Qualitative Evaluation: MAE vs PP-MAE

In this section, we complement the analysis presented in Fig. 2 of the main paper by providing qualitative examples of instances where the MAE and PP-MAE exhibit notable misalignment, deviating significantly from the primary trend observed in the graph.

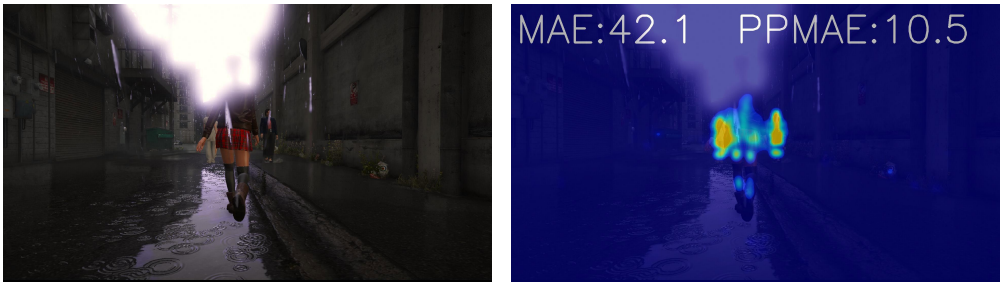
In Fig. 4, we present qualitative results illustrating that the alignment between these two metrics significantly deteriorates under rainy and dark environmental conditions. Specifically, in the first scenario, image distortion caused by raindrops leads STEERER-V to incorrectly infer volumes at a distance (as shown in Fig. 4a), while the glare from lightnings complicates the model's ability to detect and assess human figures (refer to Fig. 4b). In the context of darkness, the model can confuse environmental objects with humans, such as mistakenly identifying a tree situated between two individuals as a person (illustrated in Fig. 4c), along with other inaccuracies demonstrated in Fig. 4d and Fig. 4e.

5. Other tasks with ANTHROPOS-V

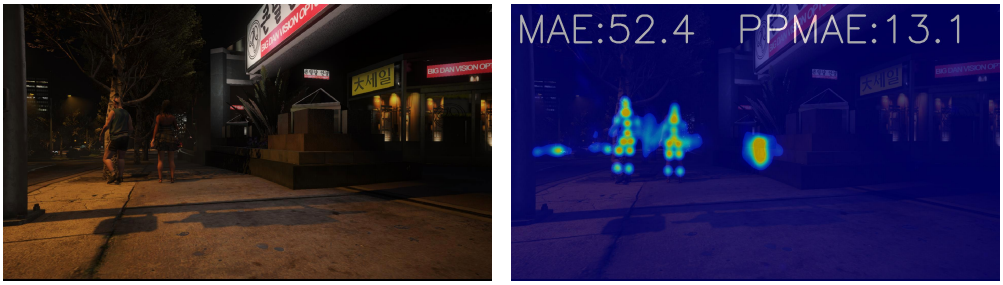
Crowd Counting	MAE	RMSE	
Bayesian+ [11]	3.50	5.87	
P2P [15]	8.38	11.7	
MAN [9]	3.54	5.88	
STEERER [6]	5.56	6.94	
HMR	**MPJPE**	**PA-MPJPE**	**PVE**
CLIFF [8]	807.6	165.9	940.8
BEDLAM-CLIFF [1]	794.5	165.7	991.0
ReFit [17]	397.2	310.2	416.3

Table 3. Results of Crowd Counting and Human Mesh Recovery on ANTHROPOS-V. MPJPE, PA-MPJPE, and PVE are measured in millimeters.

(a)



(b)



(c)

(d)

(e)

Figure 4. Qualitative evaluation including images corresponding to the points in the scatter plot in Fig. 2 of the main paper for which MAE and PP-MAE deviate from the primary trend.

We evidence that ANTHROPOS-V can further serve as a benchmark for Crowd Counting and Human Mesh Recovery (HMR), as we report in Table 3.

The low performance of HMR methods stems from the increased complexity in the lighting and weather conditions, the number of individuals in the scene, and the large number of occlusions that invalidate the person-detection step leading to inaccurate predictions. CLIFF and BEDLAM-CLIFF particularly struggle to estimate the global scale and orientation, with an MPJPE value of 807.6 mm and 794.5 mm, respectively; the error on the prediction drastically reduces to 165 mm after the Procrustes alignment. For what concerns Crowd Counting, Bayesian+ exhibits the best performance, yielding an average error of 3.5 individuals per frame and surpassing more recent methods such as STEERER.

6. Further notes on the baselines

In this section, we add some notes about the results of the HD+HMR baselines (Sec. 6.1), and the architectural adaptation of the Crowd Counting models, which we modify for the CVE task (Sec. 6.2).

6.1. About the low performance of HD+HMR baselines for CVE

Here we focus on BEDLAM-CLIFF [1].

As evidenced in Fig. 5a, the performance of the human detection model has a critical impact on the overall CVE results of the HD+HMR models. One of the failure cases originates from either a missing detection, as for the woman on the left of the pillar, or multiple predicted bounding boxes of the same instance, as for the subjects in the foreground. Also, even when the error deriving from the human detection step is driven to zero, like in the oracular experiment described in Sec. 5.1 of the main paper, BEDLAM-CLIFF underperforms when compared with STEERER-V. Indeed, occlusions, color contrast, and extreme light conditions harm the body shape regression, which in turn increases the volume estimation error, as Fig. 5b empirically confirms; for example, the woman on the left of the image is assigned with an excessively skinny SMPL mesh, and the people partially occluded by the central round terrace are approximated with an amorphous mesh.

Finally, the HD+HMR baselines yield more parameters than the baselines adapted from Crowd Counting, as shown in Table 4.

6.2. Details on the architectural adaptation of Crowd Counting baselines

We train all models on a single NVIDIA A100 GPU until convergence. Both the original codebases and the edited code leverage the PyTorch framework.

Table 4 provides further implementation details.

Model	**#Params**	**LR**
YOLOv7	165M	1×10^{-5}
CLIFF	247M	5×10^{-5}
BEDLAM-CLIFF	247M	5×10^{-5}
ReFit	240M	1×10^{-4}
MAN	40.4M	1×10^{-5}
Bayesian+	21.5M	1×10^{-5}
P2P-net	21.6M	1×10^{-5}
STEERER	64.6M	5×10^{-7}

Table 4. Details on the baselines employed for CVE. The number of parameters of the HD+HMR baselines includes the one of YOLOv7 [16], for which we report the parameters in gray on top of the table.

Bayesian+ and MAN: These models have nearly the same base architecture, so we modify them in the same way. These architectures are described by the green blocks in Fig. 6, with only MAN employing the Transformer Encoder with the Learnable Region Attention block. Bayesian+ and MAN are both Crowd Counting architectures. Hence, to adapt them to the CVE task, we define an additional branch besides the one performing counting. The orange blocks in Fig. 6 illustrate the novel branch. Since these models are trained on 512×512 image crops, Max Pooling is employed for computing volume on larger-sized images, while Pointwise Convolution compresses tensors to a single dimension. Furthermore, we alter the pre-processing pipeline to compute both counting and volume-related ground truths on which both models are supervised, i.e., the total number of persons in the frame and the total volume occupied by them, respectively.

Notice that the counting branch is necessary because we use its output, which is the estimated density map, as input for our additional volume regressive branch. For both these Bayesian-based models, the loss we use for the volume branch is the L1 loss between the regressed and the ground truth volumes. We also keep the counting losses of Bayesian+ and MAN as described in their papers.

P2P-Net: To adapt P2P-Net for the CVE task, we expand the model's capabilities to predict x and y coordinates for each identified head with a v label indicating the volume of the corresponding person. To encourage accurate predictions for the volume (v) rather than solely emphasizing x and y predictions, we introduce an additional loss component. This supplementary component is an L1 loss computed between predicted and ground truth volumes. To tune the influence of this loss, we apply a weighting coefficient λ, determined through experimentation to be optimal at the value of 1e-4.

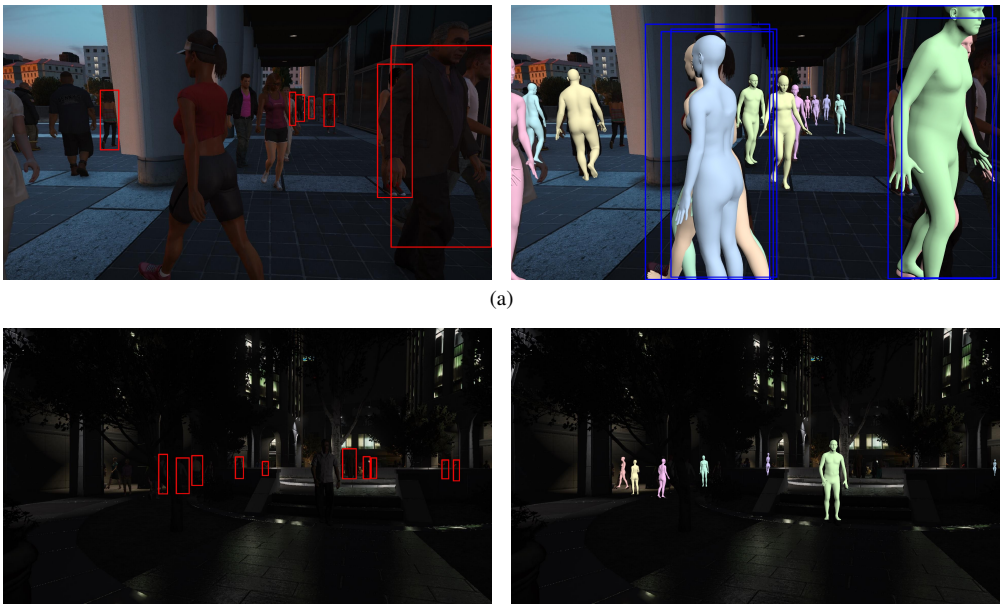


Figure 5. Qualitative results of BEDLAM-CLIFF [1] on ANTHROPOS-V when provided with the predicted bounding boxes of YOLOv7 [16] (we omit some of them for clarity). We highlight in red some of the instances that have not been detected and in blue those that have been detected multiple times.

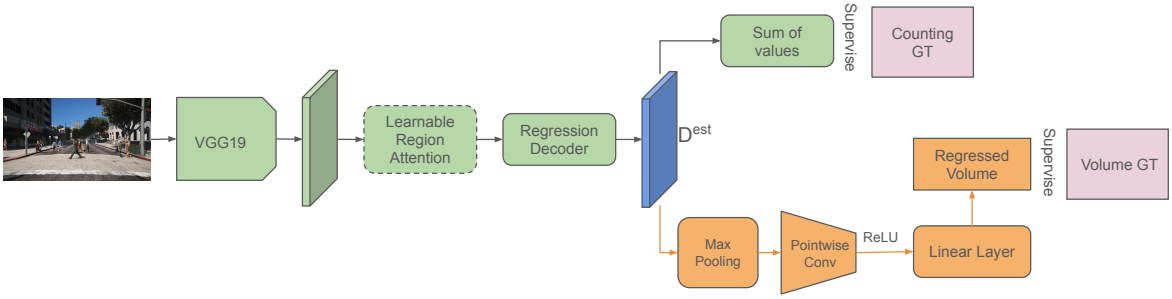


Figure 6. Bayesian+ and MAN modified architectures. The Transformer Encoder blocks with *Learnable Region Attention* are used only by MAN. The green layers are kept as they are in the original architectures. Orange layers are part of the additional branch tailored on the CVE task. The estimated density map (D^{est}) is reduced with summation to a single value and counting losses are calculated on it. D^{est} is also used as input for our additional volume-related branch to regress the volume occupied in a frame. Finally, we calculate the L1 loss between regressed and ground truth volumes.

7. Qualitative Evaluation: real-world images

We present some examples of STEERER-V's zero-shot predictions on the real-world images of CrowdHuman [13] dataset. We remark that this dataset lacks ground-truth volume annotations, which are roughly approximated by imputing the average real-world volume to each individual in the images, leveraging the statistics from [14] (cf. Sec 5.3 of the main paper). It is worth noting that such an approximation strongly assumes that people are all similar in size to the average adult and that genders are equally represented in crowds.

As Fig. 7 shows, STEERER-V reasonably estimates the volume of adults in indoor environments, large crowds, and with severe occlusions, diverging from the mean to take into account diverse builds, e.g., the men in the foreground of the first row, and uneven balance of genders, e.g., the crowds displayed in the second and third row. Our model fairly performs even with crowds of kids, totally absent from the GTA-V game. The image in the second row of Fig. 9 stresses how the model assigns greater volume to the adults in the right foreground than to the surrounding children.

STEERER-V struggles with low-quality images, such as in the example of Fig. 8.

Figure 7. Zero-shot results of STEERER-V on CrowdHuman [13] images when predicting the volume of adults. On each predicted volume map we superimpose the difference between the average real-world per-person volume and the predicted per-person one.

Figure 8. Zero-shot result of STEERER-V on a CrowdHuman [13] image. STEERER-V underestimates the total volume due to the low quality of the image, the domain gap, and the severe occlusions. On the predicted volume map we superimpose the difference between the average real-world per-person volume and the predicted per-person one.

Figure 9. Zero-shot results of STEERER-V on CrowdHuman [13] images when predicting the volume of builds that have not been seen at train time, e.g., kids. On each predicted volume map we superimpose the difference between the average real-world per-person volume and the predicted per-person one.

8. Qualitative Evaluation: additional results on ANTHROPOS-V

In Table 5, we present additional qualitative comparisons between the baseline model, STEERER, and our proposed model, STEERER-V. The comparison demonstrates that when faces are clearly visible and occlusions are minimal, the performance of both models is similar, as shown in the first and second rows of the table. However, in scenarios where occlusions occur, either due to other pedestrians or environmental elements, STEERER-V outperforms STEERER significantly, as evidenced from the third to the sixth row. In these images, it is evident that STEERER fails to attribute any volume to several individuals. Moreover, in the third row, we show that both models have learned that, from bird's-eye-view camera angles, environmental elements like trees can hide persons. Nevertheless, STEERER-V demonstrates superior robustness by not erroneously assigning any volume to the space obscured by the upper part of trees, highlighting its enhanced capability in handling such occlusions.

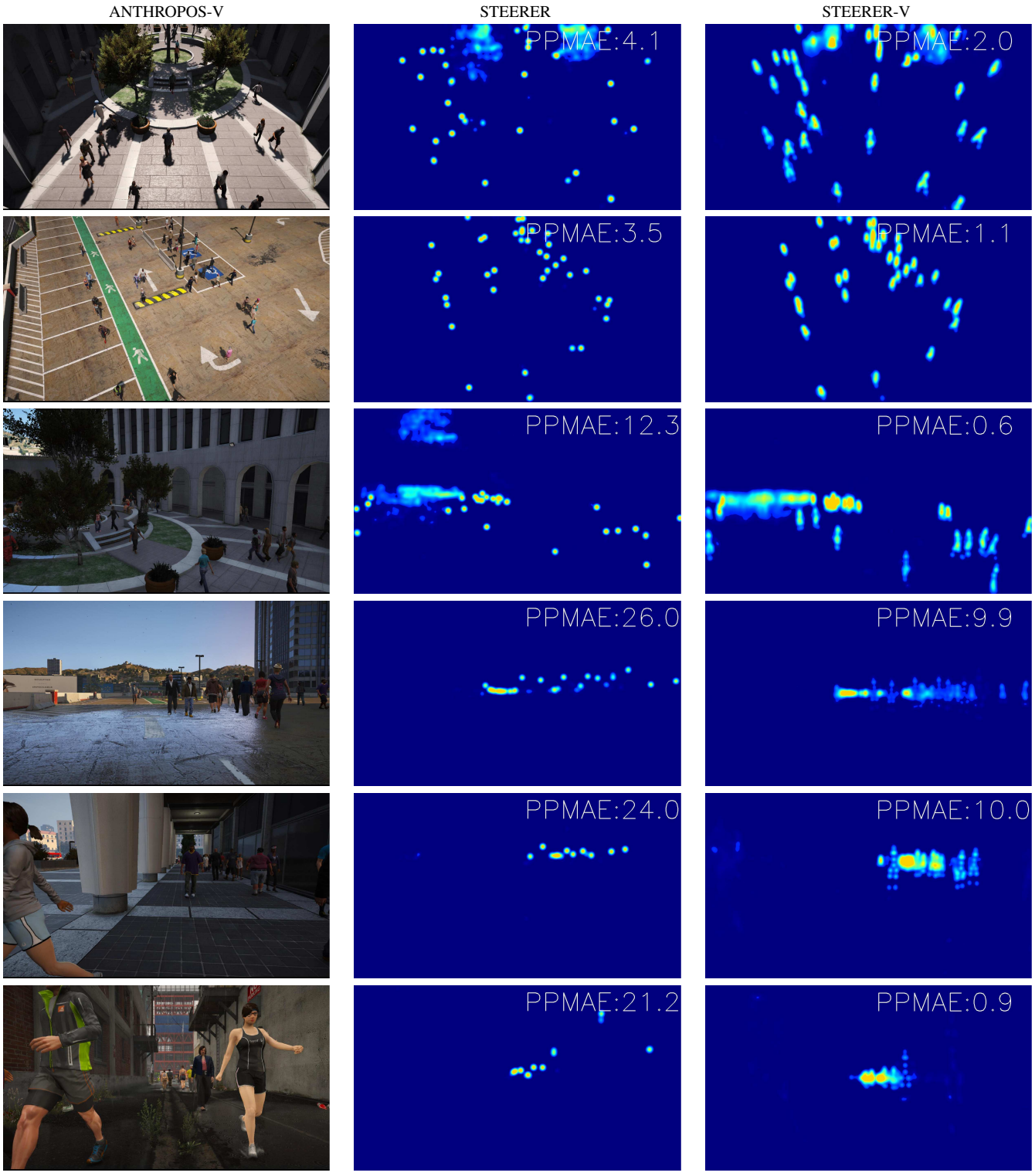


Table 5. Visualization results of STEERER and STEERER-V on ANTHROPOS-V crowded images. STEERER's density map highlights volume on head positions, while STEERER-V's density map emphasizes the volume spread on the whole body.

9. Examples of images of ANTHROPOS-V

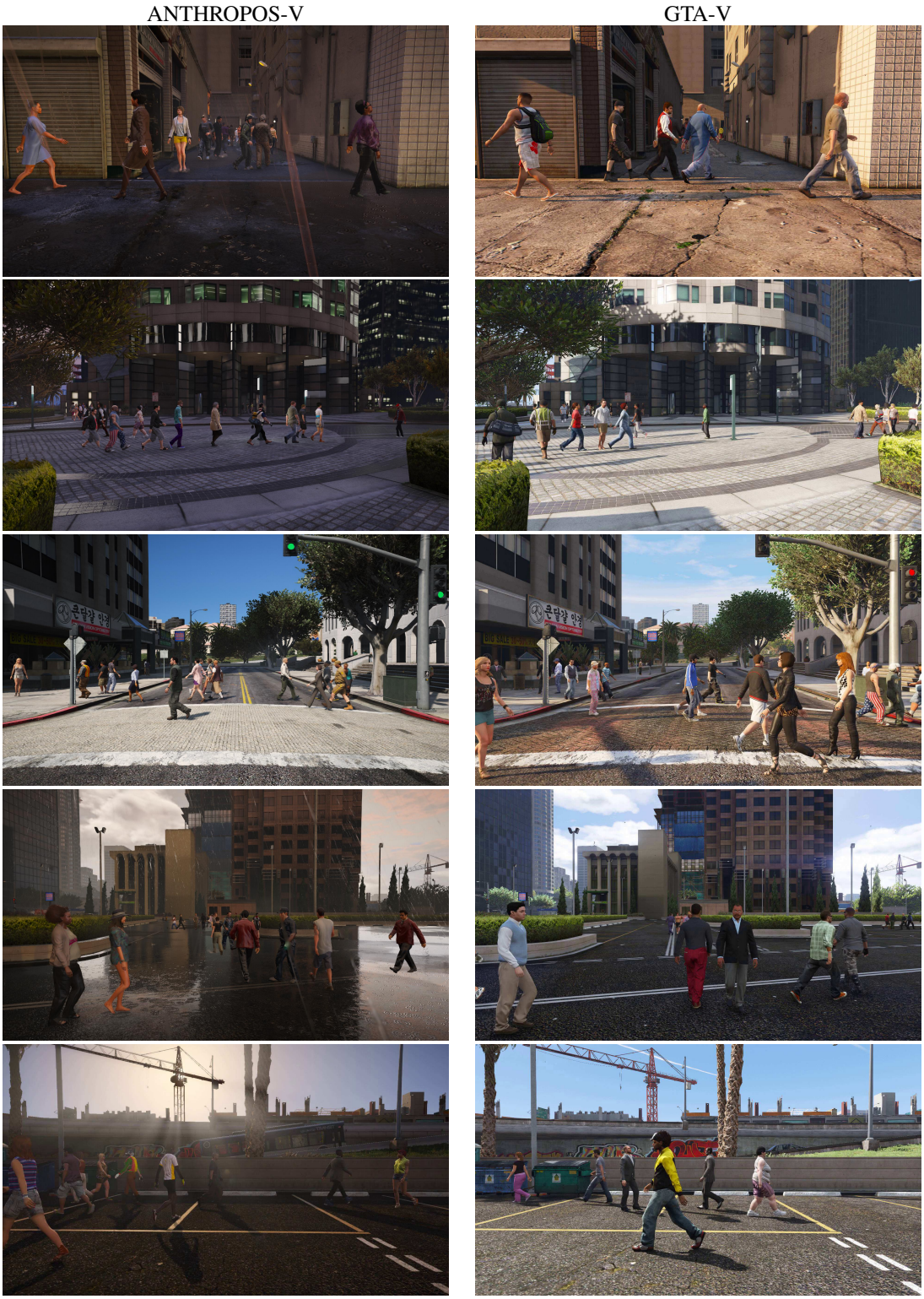


Figure 10. Examples of frames from ANTHROPOS-V (left column) and original GTA-V footages (right column), as synthesized via [4]. Images on the left present a wider variety of people's heights.

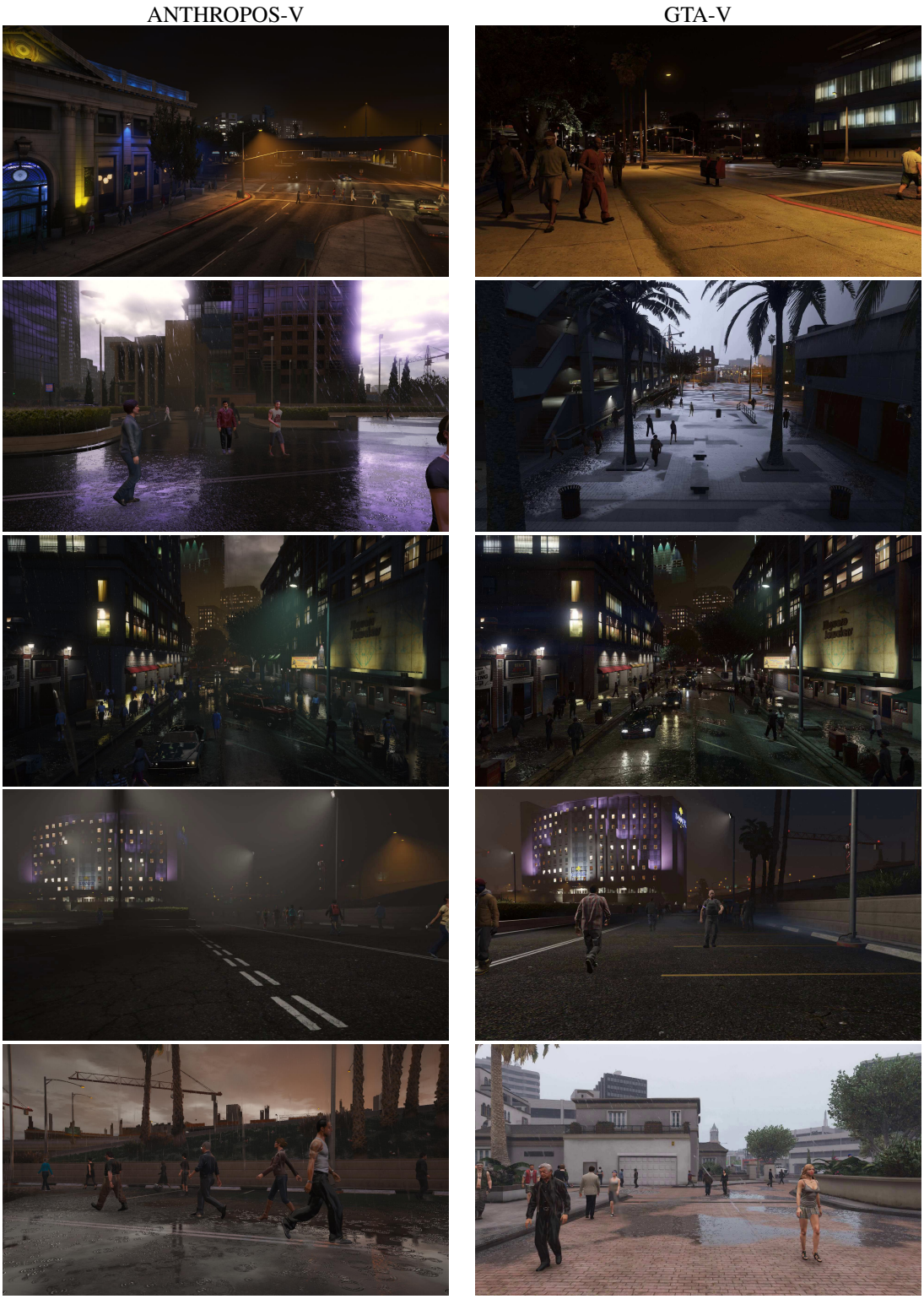


Figure 11. Examples of frames from ANTHROPOS-V (left column) and original GTA-V footages (right column), as synthesized via [4]. Images on the left present more diverse weather and lighting conditions and better details.

10. Cross Dataset Evaluation

In this section, we perform an additional experiment that assesses the performance of the HR-HMR baselines via Cross Dataset Evaluation. The next section will introduce the additional datasets we leverage for this study, while Sec. 10.2 describes the experiment's outcomes.

10.1. HMR Datasets

AGORA [12] is a synthetic image dataset with diverse adult and child characters with SMPL [10] annotation. The recent BEDLAM [1] is a comprehensive synthetic video dataset with 271 highly realistic and diverse SMPL-based characters. In contrast to our dataset, they don't target large crowds, having a limited number of people per scene ($\leq 15, 10$ in [12] and [1], respectively); hence, they are not optimal for CVE, as we show in the next section.

10.2. Cross Dataset Evaluation

To assess the capabilities and applications of ANTHROPOS-V, we conduct a cross-dataset evaluation using the latest human datasets annotated with SMPL meshes, specifically AGORA [12] and BEDLAM [1]. It is important to note that these datasets predominantly feature small groups of people, with an average of 3.66 individuals per frame in BEDLAM and 9.08 in AGORA. Since these datasets lack ground-truth volumes, we annotate the volume based on the provided meshes. In our experiment, we train STEERER-V on all three datasets and evaluate the performance on each dataset's test set. Fig. 12 displays the error rates for each test set, highlighting how they vary with the increasing number of people in a scene. As shown in Fig. 12a, models trained on datasets with smaller groups (represented by the green and blue lines) exhibit less robustness when faced with scenes containing more individuals. Conversely, Fig. 12b and Fig. 12c demonstrate that STEERER-V, when trained on our crowd dataset, maintains robustness regardless of the increasing number of people in the image.

11. Decoupling Crowd Counting from Volume Estimation

The CVE error metrics presented in this paper (MAE/PPMAE) compare per-frame ground truth volumes with model predictions. However, this error stems from two main sources: missed detections and incorrect volume estimations of individuals. To improve CVE models, both of these factors must be addressed. To assess the contribution of each error source to the overall error, we conduct an additional experiment.

We aim to design an evaluation method applicable to all the models proposed in this paper. HD+HMR models are straightforward to adapt by removing the HD component and using ground truth bounding boxes (bbox). However, density-based models require additional steps. For these models, we predict densities for the entire image, then crop the corresponding ground truth bboxes for the individuals involved, isolating their respective volumes. For all models, we consider only non-overlapping bboxes to prevent volume duplication and report the resulting PPMAE.

To avoid penalizing models for detection errors, we exclude any predicted volumes under 10 dm³ from being counted as positives. Additionally, to facilitate easier detection, we removed scenes from the ANTHROPOS-V test set that contain significant occlusions or challenging lighting conditions, creating a refined test set called S1.

Furthermore, we constructed an additional test set, S2, consisting exclusively of bird's-eye view scenes, which minimize occlusion and further simplify detection.

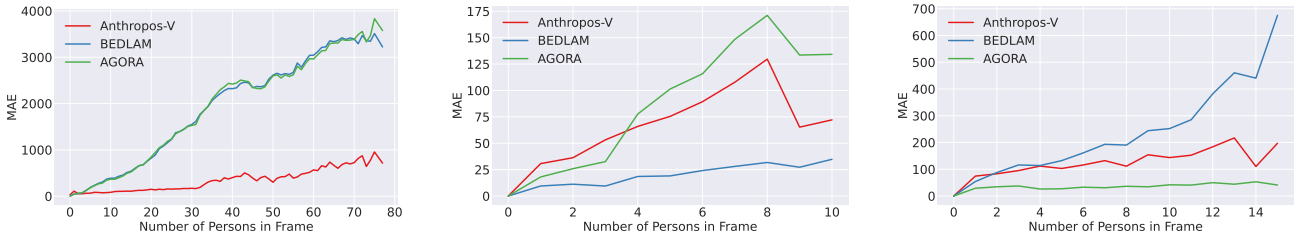
Table 6. Results on ANTHROPOS-V's S1, S2 and whole test set (FT). All the results are reported in dm^3.

Model	PPMAE (S1)	PPMAE (S2)	PPMAE (FT)
ReFit [17]	17.94	18.25	18.79
STEERER [6]	12.46	13.68	14.43
STEERER-V	6.67	3.39	6.73

In Table 6, we present the results of this experiment with simplified detection. Notably, ReFit and STEERER show error levels comparable to those on the full test set (FT), highlighting that the primary source of their error lies in evaluating individuals' volume. In contrast, STEERER-V's performance improves by 50% on the easier detection set (S2), suggesting that its error is equally divided between detection and volume estimation. Moreover, as shown in Table 1 of the main paper, when comparing both $C(I)_{B+} \times \overline{V}_D$ and ReFit with their oracular counterparts, a similar ratio emerges, further confirming that volume estimation error accounts for half of the total error.

12. From Frames to Video

Lastly, we question if leveraging temporal information can be beneficial in CVE. Specifically, we modify STEERER-V to leverage two neighboring context-frames, one before and one after the target frame. We align features from context-frames to those of the target frame using the method in [7] and feed the result into STEERER-V's decoding branch to estimate the total volume in the target frame. We use STEERER-V's pretrained weights, while the feature alignment module is trained from scratch. Despite being an initial attempt to incorporate inter-frame information, this approach proves beneficial for CVE, reducing MAE by 5.27% and PPMAE by 4.22%.



(a) Tested on ANTHROPOS-V (b) Tested on BEDLAM (c) Tested on AGORA

Figure 12. Error trends of STEERER-V with respect to the growing number of individuals.

References

[1] 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*, pages 8726–8737, 2023. 3, 5, 6, 12

[2] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

[3] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 2

[4] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Orcun Cetintas, Riccardo Gasparini, Aljoša Ošep, Simone Calderara, Laura Leal-Taixé, and Rita Cucchiara. Motsynth: How can synthetic data help pedestrian detection and tracking? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10849–10859, 2021. 10, 11

[5] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *CVPR*, 2019. 1

[6] Tao Han, Lei Bai, Lingbo Liu, and Wanli Ouyang. Steerer: Resolving scale variations for counting and localization via selective inheritance learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21848–21859, 2023. 3, 12

[7] Zhewei Huang, Ailin Huang, Xiaotao Hu, Chen Hu, Jun Xu, and Shuchang Zhou. Scale-adaptive feature aggregation for efficient space-time video super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4228–4239, 2024. 12

[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. In *European Conference on Computer Vision*, pages 590–606. Springer, 2022. 3

[9] Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, and Xiaopeng Hong. Boosting crowd counting via multifaceted attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19628–19637, 2022. 3

[10] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), oct 2015. 2, 12

[11] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6142–6151, 2019. 3

[12] 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, 2, 12

[13] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 6, 7, 8

[14] Mark P Silverman. Exact statistical distribution of the body mass index (bmi): Analysis and experimental confirmation. *Open Journal of Statistics*, 12(3), 2022. 6

[15] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3365–3374, 2021. 3

[16] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023. 5, 6

[17] Yufu Wang and Kostas Daniilidis. Refit: Recurrent fitting network for 3d human recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14644–14654, 2023. 3, 12