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

Luca Collorone*, Stefano D'Arrigo*, Massimiliano Pappa*,
Guido M. D'Amely di Melendugno, Giovanni Ficarra, Fabio Galasso
Sapienza University of Rome
{name.surname}@uniroma1.it

Abstract

*We introduce the novel task of **Crowd Volume Estimation (CVE)**, defined as the process of estimating the collective body volume of crowds using only RGB images. Besides event management and public safety, CVE can be instrumental in approximating body weight, unlocking weight-sensitive applications such as infrastructure stress assessment, and assuring even weight balance. We propose the first benchmark for CVE, comprising* ANTHROPOS-V*, a synthetic photorealistic video dataset featuring crowds in diverse urban environments. Its annotations include each person's volume, SMPL shape parameters, and keypoints. Also, we explore metrics pertinent to CVE, define baseline models adapted from Human Mesh Recovery and Crowd Counting domains, and propose a CVE-specific methodology that surpasses baselines. Although synthetic, the weights and heights of individuals are aligned with the real-world population distribution across genders, and they transfer to the downstream task of CVE from real images. Benchmark and code are available at github.com/colloroneluca/Crowd-Volume-Estimation.*

Figure 1. ANTHROPOS-V is the first dataset for the novel task of Crowd Volume Estimation. It features human crowds engaged in activities within real-world environments (*left*). Each individual in the dataset is labeled with ground-truth SMPL shape parameters (*center*) along with their volume labels. We employ novel Per-Part Volume Density Maps as a superior supervision signal for training models to address this task (*right*).

1. Introduction

Dealing with large gatherings in public spaces presents significant challenges in crowd management: *overcrowding* can jeopardize the safety, health, and comfort of individuals, while the assembly of crowds on structures not designed for high capacity poses risks of structural damage or collapse due to *overloading* [52].

Currently, the monitoring of crowds' risks based on head count [11, 50] tends to disregard potentially critical factors such as weight, occupancy, heat dissipation, and oxygen consumption, which are strongly correlated with individuals' body build [20, 52]. Additionally, these factors exhibit significant variability based on age, gender, ethnicity, and health conditions [6, 12, 19].

*Authors contributed equally.

A precise estimate of the *crowd's total volume* offers a more reliable method for detecting space underuse or overcrowding by leveraging a priori knowledge of the available in-place volume. Additionally, this approach can significantly mitigate the risks associated with overloading, as volume serves as a robust proxy for estimating weight [7, 38].

Motivated by these insights, this paper introduces the novel *Crowd Volume Estimation* (CVE) task, which aims to estimate the collective volume occupied by groups of individuals directly from single RGB images.

Recently, volume estimation (VE) has garnered significant research attention. However, previous works [15, 22, 29, 38, 39] have focused primarily on estimating the volume of single individuals in controlled environments, often relying on expensive annotations. These works rely on requirements and assumptions that render them impractical for estimating the volume of large crowds. In addition, there are no datasets tailored for CVE, as existing datasets [1, 37] featuring scenes with multiple people only encompass a limited number of individuals (≤ 15) or lack necessary volume annotations [9, 10]. We thus propose the first CVE benchmark, including baselines, a novel dataset, and metrics.
As for the baselines, we outline two research directions for crafting models for CVE: (1) grounding on methods from the Crowd Counting [13, 27, 33, 49] domain, or (2) repurposing the pipeline of Human Mesh Recovery (HMR) [1, 25, 59] combined with a human detector. In the case of (1), we ex-

tend the density-map strategy to a more specific form of supervision, the *Volume Density Maps*, tailored to regress the human volumes in an image. For these models, we further define a more fine-grained form of supervision which we dub *Per-Part Volume Density Maps* (rightmost picture in Fig. 1). This supervision allows the model to learn how to regress the volume of different human subparts (e.g., arms, chest, legs), resulting in a more accurate estimate of each individual's volume. In the case of (2), we show that HMR models with a human detection preprocessing can serve as volume estimators without any modification, as the volume estimate comes as a by-product after postprocessing the predicted meshes (middle picture in Fig. 1).

Moreover, to enable the training of CVE-specific models, we introduce "ANTHROpometrics POse Shape and Volume estimation dataset" (ANTHROPOS-V), a synthetic, large-scale, and video-realistic dataset representing large crowds in urban scenarios and reporting for the first time the annotated volume of each individual appearing in the scenes. ANTHROPOS-V is generated using the videogame engine of Grand Theft Auto V (GTA-V), which includes a large variety of realistic urban environments and a broad diversity of characters' ethnicities.

It is worth noting that acquiring real-world data with annotated anthropometric features for crowds requires gathering sensitive information from thousands of individuals in a wide range of environments, hence posing severe issues of feasibility, bias, and personal details disclosure. Thus, we collect a synthetic dataset to train models on CVE, evaluating their learned knowledge of both synthetic and real-world data.

Aiming at reducing the domain gap from synthetic to real images, we enhance the game's appearance and perform an in-depth analysis of the GTA-V default characters. Our findings reveal that the default characters exhibit a restricted and repetitive assortment of human anthropometrics, such as body sizes and heights. Therefore, we directly manipulate characters' 3D meshes to align them with the real-world human size distributions [44]. As a result, we improve both the realism of the original GTA-V scenes and their annotations, featuring virtual characters whose height and weight distributions closely follow the authentic human variations [46].

In summary, our contributions are four-fold:

- we propose the novel task of *Crowd Volume Estimation* (CVE) to regress the volume of large groups of people from RGB images;
- we release the first CVE benchmark, including metrics and baselines;
- we introduce ANTHROPOS-V, a dataset explicitly devised for CVE but also encompassing annotations for other human-centric tasks, with careful attention to mirror real-world anthropometric and gender statistics;
- we experiment with a novel volume-specific form of supervision, namely Per-Part Volume Density Maps, and use it to train our proposed model, STEERER-V, achieving superior results.

2. Related Works

In this section, we review the existing literature that relates to the proposed CVE task. We discuss studies on single-subject Volume Estimation (VE) (Sec. 2.1), Crowd-Counting (Sec. 2.2) and Human Mesh Recovery (Sec. 2.3).

2.1. Single subject VE in controlled environments

Previous literature explored volume estimation, targeting single-bodies [15, 22, 29, 38, 39] or objects [2, 5, 30, 31, 40, 61, 62] for applications in healthcare and nutrition. In particular, [29] rely on 3D scans, while [15, 30, 31, 38] exploit depth-maps or point clouds data. To deal with scale ambiguity, [5, 14, 40, 61, 62] make use of reference objects, while [2, 39] employ multiple images of the same object in different views. While all the mentioned works tightly depend on controlled environments, scans, or multiple inputs, and apply to a single subject at a time, we aim at estimating the total volume of human crowds in the wild. Notably, [22] proposed a large-scale video dataset displaying individual textured SMPL meshes [32], paired with body-part volume ground truths. However, scenes are designed by superimposing a single mesh onto 2D bedroom images, lacking realism and scale consistency between humans and the background. On the contrary, we propose a dataset of realistic scenes featuring large crowds.

2.2. Crowd Counting

Crowd Counting aims to estimate the number of people in images or videos. Typically, datasets in this domain showcase large crowds from bird-eye views [17,18,47,48,57,65]. While seminal works [4, 21, 35] cast this problem as a regression task, recent literature [26, 27, 49] address crowd counting as a localization task. These methods regress the 2D positions of the heads in the images and estimate the total number by summing the retrieved outcomes after filtering the more uncertain predictions. Density-Map-based methods [3, 13, 24, 28, 34, 41, 54, 55, 58, 64] express the ground truth density map y for an image x as a single-channel image of the same size, where each pixel is assigned 1 if it contains the center of a person's head, 0 otherwise; y is subsequently smoothed with a Gaussian filter. The Gaussian filtering operation is common in Counting tasks [23, 28, 42, 51, 63], as it allows to treat the GT density map as a continuous function [65], which, in turn, allows the end-to-end training of the network. Bayesian-based approaches differ from conventional density-based methods as they estimate density maps without supervision upon ground truth density maps. For instance, [27, 33] employ a bayesian-loss to construct a density contribution probability model starting from bare annotations. We evaluate strong Density-Map [13], Localization [49], and Bayesian models [27, 33] for the CVE task. We introduce baselines capitalizing on an adaptation of the Density Maps approach for CVE, namely Volume Density Maps: each pixel containing a person's head is assigned with the whole person's body volume instead of 1.

2.3. Human Mesh Recovery for Few Individuals

Human Mesh Recovery (HMR) regresses the human 3D shape and pose from single RGB images. CLIFF [25] pairs cropped image features with their bounding box information, enabling the accurate prediction of global rotations. In BEDLAM-CLIFF [1], the authors train CLIFF on their dataset and improve its performance. ReFit [59] exploits a recurrent updater that iteratively adjusts a parametric human model to align with image features. The recent TokenHMR [8] uses a tokenized representation of the human pose and reformulates the problem as a token prediction. DPMesh [66] leverages a diffusion model to meliorate robustness to occlusions. Crowd3DNet [60] focuses on mesh reconstruction of people within crowds in wide-field images, though tightly assuming the existence of a common plane where all the actors lie; such assumption does not hold for the complex scenes of ANTHROPOS-V. Similarly, [16] exploits pseudo-GT to model the relations and interactions of the individuals and improve pose and localization estimation; however, this work do not focus on human shape, as no shape-related metric is employed. Contrarily, CVE requires precise volume/shape GT for correct computation (Sec. 3.2). Thus, we repurpose [1, 25, 59] for CVE, pairing them with a human detector.

3. Measuring Crowds Volumes

In this section, we formalize CVE (cf. Sec. 3.1) and define the metrics for the novel CVE benchmark (cf. Sec. 3.2).

3.1. Problem Formalization

We define **Crowd Volume Estimation** as the task of estimating the undergarment total body volume occupied by human bodies in a given scene. While CVE can be applied to videos (that we make available in ANTHROPOS-V), we define the CVE task to be benchmarked per frame.

Let I be an image and V_{tot} the label of the actual total volume of human bodies represented in I. We define the objective of CVE as $\min_{\theta} ||V_{tot} - M_\theta(I)||$, where M_θ is a crowd volume estimation function parameterized by θ. This

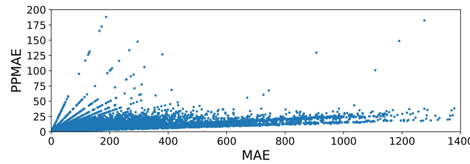


Figure 2. STEERER-V's PP-MAE/MAE computed on ANTHROPOS-V test set.

definition is intentionally general and designed to be independent of any specific methodology. Indeed, this formulation enables its application as an objective for both Crowd Counting and HMR models, as well as for our proposed method.

3.2. Proposed Metrics

Leveraging earlier research in Crowd Counting [13, 18, 57, 65], we propose to measure the volume estimation error with the following suite of metrics:
Mean Absolute Error (MAE), as a standard measure to assess the quality of the estimations. Given a set of images $\{I_k\}$, we indicate with $\{V_k\}$ the total volume associated to each image and with $\{\hat{V}_k\}$ the estimated one. Hence:

$$\text{MAE} = \frac{1}{K} \sum_{k=1}^{K} |V_k - \hat{V}_k| \tag{1}$$

This measure estimates the accuracy of the predictions on the whole test set as K represents the total count of the images belonging to the test set.
Per-Person Mean Absolute Error (PP-MAE), to measure the average error for each individual. Its formulation can be expressed as:

$$\text{PP-MAE} = \frac{1}{K} \sum_{k=1}^{K} \frac{|V_k - \hat{V}_k|}{n_k} \tag{2}$$

where n_k is the number of persons in the k-th image. PP-MAE highlights the estimator's mean error per person, enabling evaluation across scenes and datasets with different numbers of individuals. This is advantageous in CVE, as crowd sizes vary widely among frames. It is worth noting that PP-MAE is related to the Normalized Absolute Error metric used in Crowd Counting [13], which normalizes the mean counting error by the total number of persons.
Although PP-MAE and MAE appear to be closely related, Fig. 2 confirms the absence of a direct correlation between the two metrics. The lines composed of aligned points in Fig. 2 correspond to frames with a fixed human count (as their slope MAE/PP-MAE equals the number of individuals). Thus, comparing MAE vs. PP-MAE facilitates identifying the frames whose estimation error marginally depends on the number of individuals and more likely stems from other latent variables (*e.g.* camera position, weather and lighting condition).

4. Crowd Volume Estimation

In this section, we outline how we repurpose Human Mesh Recovery (HMR) (Sec. 4.1) and adapt Crowd Counting (Sec. 4.2) baselines for our task. In Sec. 4.3, we describe the intuitions and methodologies of our proposed approach.

4.1. From HMR to CVE

We adapt CLIFF [25], BEDLAM-CLIFF [1], and ReFit [59] to the task of CVE according to the following pipeline: (1) identifying human occurrences using a human detector (HD) model, (2) determining the mesh for each individual in the scene, and (3) calculating the volume of each mesh. Finally, the total crowd volume is obtained by aggregating the individual volumes. We dub these baselines as HD+HMR. A shortcoming of this approach is that these methods rely upon an upstream human detector, which can fail when multiple human instances populate an image, as in the case of crowds. To marginalize this issue, we also consider an oracular baseline that replaces the predicted bounding-box locations of humans in the scene with the ground truth ones.

4.2. From Crowd Counting to CVE

To assess whether CVE can be naively solved without the adoption of any specific strategy, we set a baseline whose volume estimation stems from $C_{B+}(I) \times \bar{V}_{\mathcal{D}}$, where C_{B+} is a Crowd Counting model[1], I is the input image, and $\bar{V}_{\mathcal{D}}$ is the average per-person volume in the dataset $\mathcal{D}$. As a statistical reference, we further experiment with an oracular version of this baseline that replaces $C_{B+}(I)$ with the ground-truth count of image I, namely Oracular $C(I)\times\bar{V}_{\mathcal{D}}$.

Additionally, we adapt relevant *Localization*, *Bayesian*, and *Density Map* approaches from Crowd Counting (cf. Sec. 2.2). As for the *Localization* approach, we select P2P-Net [49]. For CVE purposes, we adjust its architecture to predict an array of (2+1) scalars, where the additional coordinate represents the volume of the target person.

As *Bayesian* approaches we consider Bayesian+ [33] and MAN [27]. We adapt them for CVE by appending an additional branch that takes the estimated density map as input and regresses the total volume in the input frame (cf. Sec. 6.2 and Fig. 6 in the Supplementary Material).

For the *Density Map* method, we adopt the recent STEERER [13]. Our adaptation preserves the original network architecture while modifying the model's supervision technique: instead of using conventional counting density maps that label a pixel representing a person's head with the value 1, we use *Volume Density Maps*, where we annotate the pixel with the person's total volume. This Volume Density Map is then smoothed using a Gaussian filter.

[1] We use Bayesian+ [33]. In the Supplementary Material, we demonstrate that, for counting purposes, it performs best on ANTHROPOS-V.

4.3. Per-part Volume Density Maps

In our proposed approach, we leverage ANTHROPOS-V per-part volume annotations, discussed in Sec. 6.3. Driven by the insight that volume is distributed throughout the human body, we enhance the proposed Volume Density Maps approach to incorporate this concept. Specifically, since ANTHROPOS-V provides fine-grained annotations of body parts volumes of each character, we introduce *Per-Part Volume Density Maps*, where specific keypoints of each person are assigned a portion of the total body volume. For instance, each of the five torso keypoints will be attributed with $\frac{1}{5}$ of the torso-only volume. After smoothing this local map, the volume is distributed over the interested body parts (see the second column of Table 3 for visualization). We train a STEERER-like model from scratch with these annotations and dub this model as STEERER-V.

5. Experiments

In this section, we evaluate all methods' performance on ANTHROPOS-V quantitatively and qualitatively (Secs. 5.1 and 5.2). Sec. 5.3 provides the results of our best model on real-world datasets.

5.1. Experimental Results

Table 1 reports results on the test set of ANTHROPOS-V for the CVE task. All the baselines are trained on ANTHROPOS-V. The HD+HMR baselines' human detectors are YOLOv7 [56] instances fine-tuned on our dataset.

HD-HMR methods do not perform well. The upper part of Table 1 shows that HD+HMR methods report suboptimal performance in CVE tasks. These approaches are limited by the heterogeneous scales of individuals within crowd scenes and the limitations of the HD, whose accuracy is significantly marred by severe occlusions and challenging environmental conditions. Note that HD not only fails to generate a bounding box for some individuals, but it can also propose multiple bounding boxes for the same person, resulting in redundant volume estimations for the same individual. Replacing the HD with an oracle that provides GT bounding boxes yields a marked reduction in volume estimation error, still reporting a rather large PPMAE with respect to the Crowd Counting adapted baselines. This is probably due to the elevated number of occlusions in ANTHROPOS-V, which hampers the exact body shape reconstructions (cf. Sec. 6.1 of Sup. Mat.). Our proposed STEERER-V demonstrates superior performance, surpassing all the oracle-enhanced HMR approaches.

Density maps help in CVE. STEERER demonstrates superior performance among the methodologies adapted from Crowd Counting and trained on Volume Density Maps (second block in Table 1). This result suggests that Bayesian [27, 33] and localization [49] techniques exhibit

	Model	Oracle	MAE	PPMAE	Inf. time
HMR	CLIFF [25]		673.7	21.41	145.7
	BEDLAM-CLIFF [1]		656.4	21.17	137.9
	ReFit [59]		595.2	18.79	170.0
	CLIFF [25]	✓	370.2	12.89	56.68
	BEDLAM-CLIFF [1]	✓	364.7	12.15	49.73
	ReFit [59]	✓	346.8	11.31	108.1
Counting	Bayesian+ [33]		578.09	17.31	**37.50**
	P2P [49]		590.91	17.07	61.16
	MAN [27]		557.90	17.03	81.84
	STEERER [13]		506.94	14.43	105.1
	$C_{B+}(I) \times \bar{V}_{\mathcal{D}}$		507.97	14.39	**37.50**
	Oracular $C(I) \times \bar{V}_{\mathcal{D}}$	✓	191.50	5.32	-
	STEERER-V [13]		**205.59**	**6.73**	105.1

Table 1. Results on ANTHROPOS-V, reported in dm^3. Inference time (ms) is measured on an NVIDIA A100. Methods are divided into HD+HMR, Crowd Counting, and our proposed approach. $C_{B+}(I)$ refers to the headcount given by [33], while $\bar{V}_{\mathcal{D}}$ is the average human volume. Grayed-out lines use oracular information and should not be directly compared with the other results.

	$C_{B+}(I) \times \bar{V}_{\mathcal{D}}$	Refit	B-CLIFF	CLIFF	STEERER	STEERER-V
3DPW	71.3/43	125/75	64.0/40	69.8/43	102/89	40.4/25
CH	-	-30.0	-10.1	+1.00	-7.30	-3.40

Table 2. Evaluation on 3DPW and CrowdHumans (CH). Reported metrics are MAE/PP-MAE (3DPW) and the difference between the average real-world per-person volume and the predicted per-person one (CH). B-CLIFF stands for BEDLAM-CLIFF.

suboptimal efficacy in CVE compared to architectures purely based on density prediction. Indeed, STEERER-V, a model built up from STEERER that benefits from the proposed Per-Part Volume Density Maps during training (cf. Sec. 4.3), reports the best performance. This superiority is attributed to its capacity for fine-grained predictions, enabling effective management of significant occlusions inherent in crowded scenarios. Quantitatively, STEERER-V reports a minimal average error of 6.73 dm^3 per individual, representing a 53.36% improvement over the most effective Crowd Counting adapted model, STEERER.

Crowd Counting is not enough for CVE. Additionally, Table 1 presents the results of the $C_{B+}(I) \times \bar{V}_{\mathcal{D}}$ approach (Sec. 4.2). This method underperforms when compared to STEERER-V and its oracular counterpart. This is due to the compounded error arising from substituting individual body volume estimations with the average volume, $\bar{V}_{\mathcal{D}}$, as well as the inherent detection inaccuracies of the counting model, C_{B+}. This comparison underscores the significant performance degradation that would result *in practice* from naively applying a Crowd Counting strategy to CVE, highlighting the necessity for a specialized approach designed specifically for VE.

When considering Oracular $C(I) \times \bar{V}_{\mathcal{D}}$, which mimics a perfect human detector, something unattainable in practical applications, the error is reduced. This comparison emphasizes the magnitude of the error introduced by the imperfect detection carried out by $C_{B+}(I) \times \bar{V}_{\mathcal{D}}$.

It is worth noting that although STEERER-V does not leverage any privileged information and consequently exhibits imperfect detection, it is comparable with Oracular $C(I) \times \bar{V}_{\mathcal{D}}$. This indicates that STEERER-V compensates for its detection inaccuracies achieving a robust per-person volume estimation, making it the best candidate for practical CVE application. This quality primarily originates from STEERER-V's training strategy, which integrates body part detection, mitigating false negatives caused by occlusions, with expert knowledge of the volume contribution of each body segment. An additional experiment where we separately assess the contributions of the volume estimation and detection errors to the total error is available in Supplementary Material's Sec. 11.

5.2. Qualitative Evaluation

Table 3 presents the qualitative results of our proposed method, STEERER-V, alongside the Per-Part Volume Density Map, which serves as its training supervision. Furthermore, we provide qualitative results of STEERER and BEDLAM-CLIFF. STEERER-V stands out as the top performer because of its robustness to occlusion and its capacity to generalize. The first row demonstrates that both STEERER and STEERER-V perform well when heads are visible and occlusions are minimal. However, STEERER tends to hallucinate volume along the branches of trees, probably because the model learned that such an object may hide human heads. Contrarily, STEERER-V, designed to distribute volume across the entire body, does not suffer from this side effect, as it does not detect bodies in such scenarios. As occlusions intensify, particularly with multiple people overlapping at a distance (second and third rows), the performance gap between STEERER and STEERER-V becomes more pronounced, with STEERER-V being notably better. In the case of the dark image in the fourth row, STEERER fails to recognize the volume of the person in the foreground because their head merges with the background, while STEERER-V focuses on visible body parts, such as arms or legs, thus reducing the error. Additional qualitative results are available in the Supplementary Materials.

5.3. From ANTHROPOS-V to real images

We assess the transferability of models trained on ANTHROPOS-V to real imagery for CVE. Given the absence of suitable real-world datasets with volume annotations for crowds, we employ a bifurcated evaluation approach: we use crowd-centric real-world datasets, such as CrowdHumans [45], which lack volume annotations, and mesh-based real-world datasets, such as 3DPW [53], which allow ground-truth volume computation but do not feature crowds. For CrowdHumans [45], we address the lack

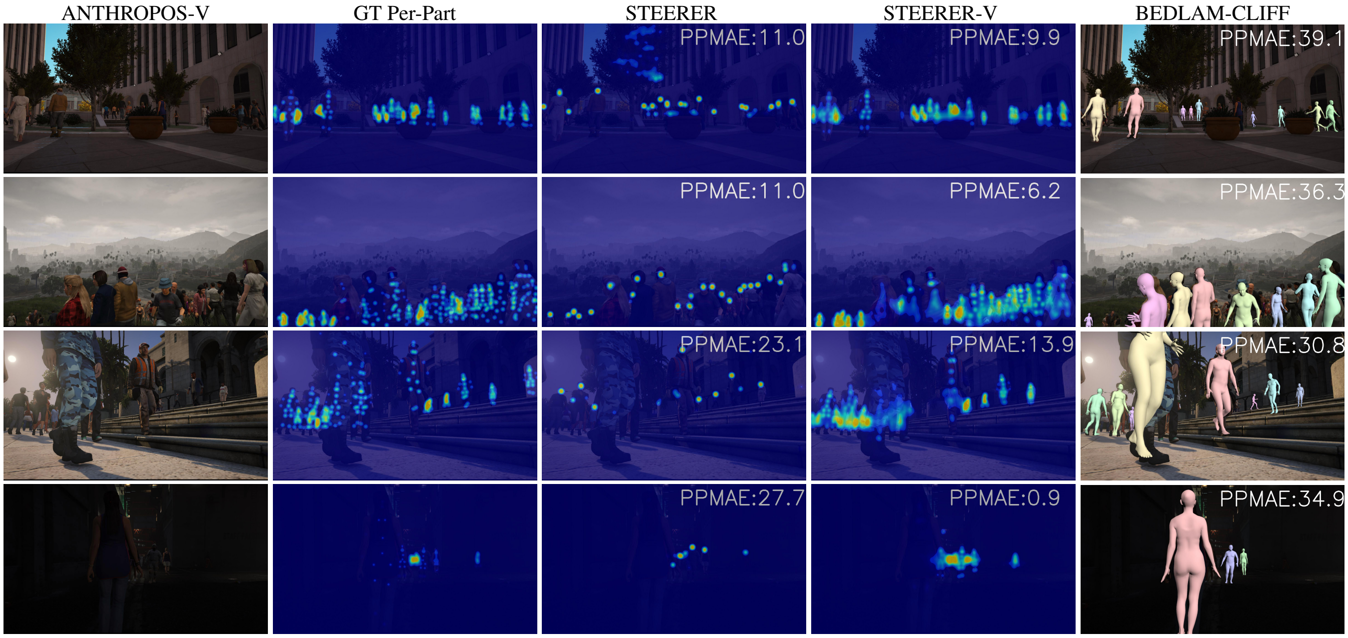


Table 3. Visual results of baseline models and STEERER-V on ANTHROPOS-V, along with the Ground Truth Per-Part-Volume Density Maps (GT Per Part). STEERER's concentrates the volume on the heads, whereas STEERER-V distributes it across the entire body.

of volume labels by imputing the average real-world volume [46] for each individual in the images. We compare these estimates with each model's predictions (Table 2). This experiment assesses the alignment of each model's predictions with expected crowd volumes, with STEERER-V and CLIFF being the most aligned. STEERER-V underestimates the expected volume by 3.40 dm^3 per person, while CLIFF overestimates it by 1.00 dm^3. Qualitative results on CrowdHumans are provided in the Supplementary Materials. For 3DPW [53], we compare each model's predictions against ground-truth mesh volumes. However, several 3DPW images include unannotated persons, such as cameramen or unscripted passers-by. Since no ground-truth is available for these individuals, we manually excluded these images from our test set, reducing the original test set to 6989 images. Results indicate that STEERER-V trained on ANTHROPOS-V outperforms all baseline models (Table 2), with MAE and PPMAE registering at 40.40 and 25.28, respectively. Additionally, we evaluated STEERER-V trained on datasets from [1] and [37] on 3DPW. In this scenario, STEERER-V continues to showcase superior results, with its counterparts presenting increased MAE and PPMAE to (59.72, 37.43) and (44.47, 29.15), respectively.

6. The ANTHROPOS-V dataset

Here we describe the generation of the proposed ANTHROPOS-V (Sec. 6.1). We detail how we align in-game meshes to the real-world statistics (Sec. 6.2) and how we obtain SMPL meshes (Sec. 6.3). We also comment on ANTHROPOS-V statistics and annotations (Sec. 6.4).

6.1. Dataset Generation

We construct ANTHROPOS-V exploiting the tools introduced in [9, 10], which, leveraging the game engine from Grand Theft Auto V (GTA-V), allow us to create densely crowded scenes within photorealistic environments. GTA-V provides several 3D urban settings, with different weather and lighting conditions during day and night, and a broad array of characters with diverse appearances, as depicted in Fig. 3. In addition, differently from previous GTA-based datasets [9, 10], to achieve a higher degree of photorealism, we use a professionally designed mod [43] that enhances the game graphics and improves the behavior and interaction among characters. Moreover, it offers additional atmospherical conditions and improves the physics in the scenes.

6.2. Alignment to real-world body-types

The original GTA-V meshes exhibit a narrow range of variations in anthropometric features, with utterly repetitive heights and volumes and a noticeable imbalance in gender representation. To address this limitation, we carefully revise the in-game meshes and code and generate a distribution of individuals that closely mirrors the real-world one [44] concerning height, volume, and gender.

To achieve this purpose, first, we conduct an in-depth statistical analysis of the distribution of the characters' anthro-

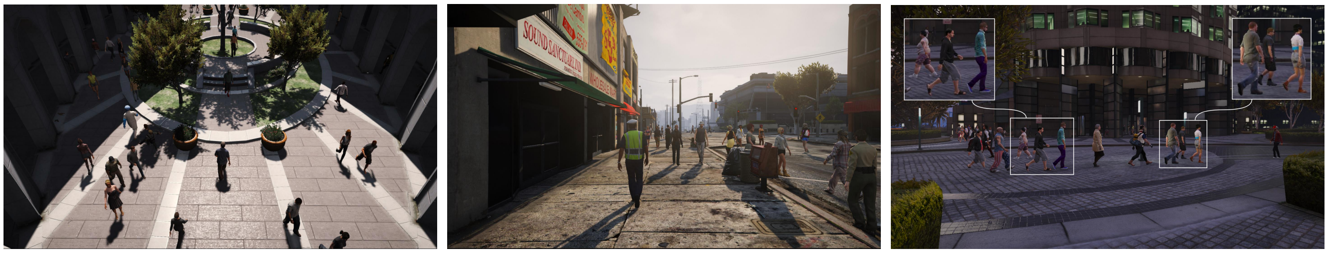
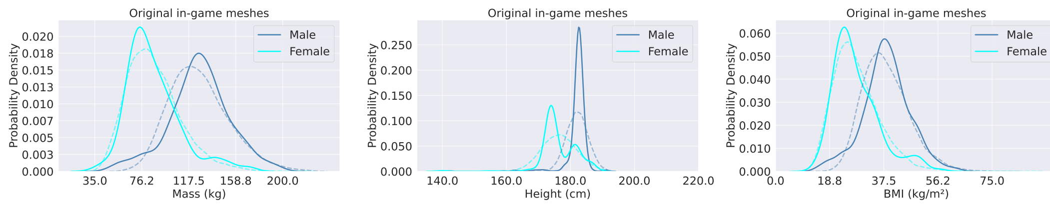
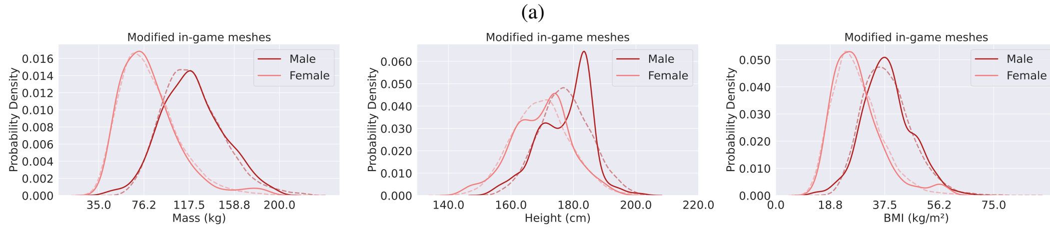


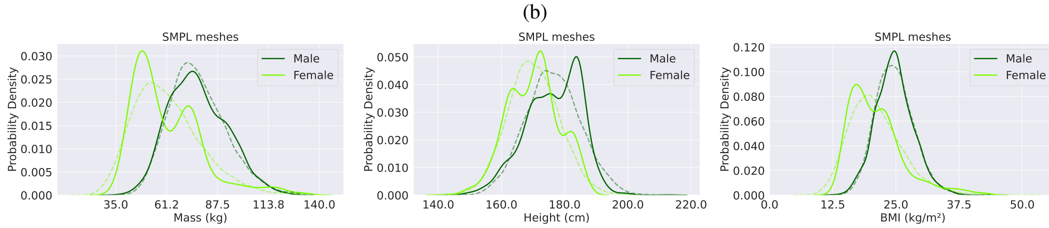
Figure 3. Examples from ANTHROPOS-V, showcasing several lighting and weather conditions, camera angles, and a variety of physiques. The crops in the zoomed boxes depict persons with differences in statures and body shapes.



(a)



(b)



(c)

Figure 4. Statistical analysis of the distributions of mass, height, and *Body Mass Index* (BMI) of the individuals in ANTHROPOS-V. Solid curves depict the empirical distributions, while dashed curves refer to the theoretically expected ones [46]. (4a) Distribution of the body features of the characters' meshes in GTA-V without any manipulation. (4b) Distribution of the body features of the characters' meshes in GTA-V after applying some geometrical transformations. (4c) Distribution of the body features of the resulting fitted SMPL meshes.

pometrics in GTA-V. We consider the mass[2], height, and *Body Mass Index* (BMI) of the male and female characters in the game (from now on referred to as "in-game meshes"). We estimate the mass by multiplying the body volume by the average body density (1000 kg/m^3 as in [7, 38]).

As theoretically proven by [46], such body features can be represented with random variables M, H, B that follow a log-normal distribution $\Lambda(\mu, \sigma^2)$:

$$M \sim \Lambda(\mu_M, \sigma_M^2), \quad H \sim \Lambda(\mu_H, \sigma_H^2), \quad B = \frac{M}{H^2} \tag{3}$$

Fig. 4 shows the empirical distributions (solid lines) as opposed to the expected distributions (dashed lines).

[2]medical literature refers to the body mass as "weight", which in physics refers to another quantity; we stick with the physics definition.

The body features of the original in-game meshes do not adhere to the theoretically expected ones, especially for the height that varies in a narrow range around the mean, as evident in the middle plot in Fig. 4a. To mitigate such mismatch and increase the variance, we scale the in-game meshes along the three axes with scaling factors α, β, γ that we independently sample from truncated normal distributions; we carefully choose the hyperparameters for this step to avoid unfeasible and unnatural bodies and to end up with meshes that appear realistic (qualitative results of the scaling are reported in the Supplementary Material). The anthropometrics of the resulting meshes follow a distribution that improves the approximation (Fig. 4b). Quantitatively, the Kullback-Leibler divergence between the empirical and the expected distributions, averaged across genders, decreases by 27.9%, 63.3%, and 19.8% for mass, height, and BMI, respectively. The SMPL fitting process (Sec. 6.3) disregards the clothing, thereby producing meshes that more closely match the real-world distributions of height [44] and BMI [36] (Fig. 4c). As a final remark, the BMI of the SMPL meshes in ANTHROPOS-V ranges in $[10, 50]$ kg/m^2, representing also underweight and obese individuals.

6.3. SMPL Fitting

To label each character with accurate ground truth volume, we employ a technique akin to the one described in [37]. The fitting procedure ensures that the SMPL mesh tightly conforms to the character mesh's uncovered body parts while allowing a looser fit on clothed parts. Details about this process are described in the Supplementary Materials. We report that our SMPL meshes have an average per skin vertex error of 7.32 mm and a penetration error of 10 mm for clothed vertices, where a looser fit is desired. This measure indicates how much these vertices extend beyond the GTA-V mesh. Finally, we use the obtained meshes to compute ground-truth volume labels for each character. Notably, besides offering labels for the total body volume, ANTHROPOS-V includes annotations for the volume of individual body parts obtained by slicing the SMPL meshes. We divide the estimated meshes into nine sections: head, torso, thighs, left and right arms, forearms, and calves. We then calculate the volume of each of these parts separately.

6.4. Dataset Statistics

ANTHROPOS-V features 768 FHD videos with annotated volumes, SMPL shape parameters, keypoints, and camera parameters and position. Videos are recorded at 30 fps and display crowds moving in diverse urban scenarios. ANTHROPOS-V features 701 distinct characters, each with a variable number of outfits, resulting in over 3k unique appearances, interacting in 384 diverse scenarios with different camera angles and weather conditions. To propose a fair split, we divide characters into three disjoint sets of 495, 64, and 142 that we distribute in different train, validation, and test videos, respectively. Within crowded scenes, characters engage with each other and with the environment, undertaking interrelated actions. For instance, they avoid collisions and form queues to navigate stairs or enter confined areas.

7. Limitations and Future Works

As the first endeavor to establish a benchmark for Crowd Volume Estimation (CVE), our work lays the groundwork for this emerging field. However, we acknowledge some aspects of our work that present opportunities for future refinement.

We introduced ANTHROPOS-V aiming to bridge the gap between synthetic and real-world data. While testing the transferability of the learned knowledge on real images without fine-grained and precise volume annotations may suffice to make an initial point on the validity of the dataset, future work should embark on acquiring detailed volume estimates of real images. Moreover, it may pursue even larger crowds, increasingly complex and diverse interactions, and estimates of objects other than people (e.g. backpacks, bags, etc.). The current output of our model provides a single per-frame number representing the total crowd volume. While suitable for many applications, this approach encourages exploration into more granular spatial analyses that could further benefit fields such as civil engineering, where detailed volume distribution information might be valuable.

Finally, we acknowledge that the ethical implications of CVE from images present complex challenges. Primary among these is the privacy issue in public spaces, which intersects with concerns about data security and the potential for misuse, as the underlying data could be adapted for unintended surveillance purposes. Furthermore, bias in volume estimates due to potential underrepresentation in training data could lead to discriminatory applications. As CVE technology evolves, these ethical considerations underscore the critical need for robust guidelines and transparent deployment protocols to ensure that the benefits of CVE can be realized while safeguarding individual rights.

8. Conclusion

In this study, we have established the first benchmark for Crowd Volume Estimation. We introduced relevant metrics and developed a dataset specifically designed for this task, focusing on human crowds in real-world-like environments. Additionally, we evaluated baseline and oracular models adapted from Crowd Counting and Human Mesh Recovery domains. Furthermore, we proposed a novel supervision approach called Per-Part Volume Density Maps, which we utilized to train STEERER-V, achieving superior results. Given the challenges in gathering real-world datasets for CVE, we anticipate that introducing this new task and benchmark will ignite interest in the research community and inspire future endeavors in the field.

Acknowledgements

We acknowledge financial support from the PNRR MUR project PE0000013-FAIR and from the Sapienza grant RG123188B3EF6A80 (CENTS). Also, we acknowledge WSense and Chiara Petrioli for partially funding this work. Finally, we extend our gratitude to Matteo Fabbri for providing the materials essential for the initial setup of the dataset generation. This work has been carried out while Stefano D'Arrigo and Massimiliano Pappa were enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome.

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. 1, 3, 4, 5, 6

[2] Nándor Bándi, Rudolf-Bálint Tunyogi, Zoltán Szabó, Eszter Farkas, and Csaba Sulyok. Image-based volume estimation using stereo vision. In *2020 IEEE 18th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 000055–000060, 2020. 2

[3] Jian Cheng, Haipeng Xiong, Zhiguo Cao, and Hao Lu. Decoupled two-stage crowd counting and beyond. *IEEE Transactions on Image Processing*, 30:2862–2875, 2021. 2

[4] Siu-Yeung Cho, Tommy WS Chow, and Chi-Tat Leung. A neural-based crowd estimation by hybrid global learning algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 29(4):535–541, 1999. 2

[5] Joachim Dehais, Marios Anthimopoulos, Sergey Shevchik, and Stavroula Mougiakakou. Two-view 3d reconstruction for food volume estimation. *IEEE Transactions on Multimedia*, 19(5):1090–1099, 2017. 2

[6] Paul Deurenberg, Mabel Yap, and Wija A Van Staveren. Body mass index and percent body fat: a meta analysis among different ethnic groups. *International journal of obesity*, 22(12):1164–1171, 1998. 1

[7] John VGA Durnin and JVGA Womersley. Body fat assessed from total body density and its estimation from skinfold thickness: measurements on 481 men and women aged from 16 to 72 years. *British journal of nutrition*, 32(1):77–97, 1974. 1, 7

[8] 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*, pages 1323–1333, 2024. 3

[9] 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. 1, 6

[10] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. In *Proceedings of the European conference on computer vision (ECCV)*, pages 430–446, 2018. 1, 6

[11] Miguel Fiandeiro, Thanh Thi Nguyen, Hanting Wong, and Edbert B Hsu. Modernized crowd counting strategies for mass gatherings—a review. *Journal of Acute Medicine*, 13(1):4, 2023. 1

[12] Dympna Gallagher, Marjolein Visser, Dennis Sepúlveda, Richard N. Pierson, Tamara Harris, and Steven B. Heymsfield. How Useful Is Body Mass Index for Comparison of Body Fatness across Age, Sex, and Ethnic Groups? *American Journal of Epidemiology*, 143(3):228–239, 02 1996. 1

[13] 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. 1, 2, 3, 4, 5

[14] Hamid Hassannejad, Guido Matrella, Paolo Ciampolini, Ilaria De Munari, Monica Mordonini, and Stefano Cagnoni. A new approach to image-based estimation of food volume. *Algorithms*, 10(2):66, 2017. 2

[15] Pengpeng Hu, Xinxin Dai, Ran Zhao, He Wang, Yingliang Ma, and Adrian Munteanu. Point2partvolume: Human body volume estimation from a single depth image. *IEEE Transactions on Instrumentation and Measurement*, 72:1–12, 2023. 1, 2

[16] Buzhen Huang, Jingyi Ju, Zhihao Li, and Yangang Wang. Reconstructing groups of people with hypergraph relational reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14873–14883, 2023. 3

[17] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2547–2554, 2013. 2

[18] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2, 3

[19] Kristi R Jenkins, Nancy H Fultz, Stephanie J Fonda, and Linda A Wray. Patterns of body weight in middle-aged and older americans, by gender and race, 1993–2000. *Sozial-und Präventivmedizin/Social and Preventive Medicine*, 48:257–268, 2003. 1

[20] CT Kappagoda, RJ Linden, and JP Newell. A comparison of the oxygen consumption/body weight relationship obtained during submaximal exercise on a bicycle ergometer and on a treadmill. *Quarterly Journal of Experimental Physiology and Cognate Medical Sciences: Translation and Integration*, 64(3):205–215, 1979. 1

[21] Dan Kong, Douglas Gray, and Hai Tao. A viewpoint invariant approach for crowd counting. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 1187–1190. IEEE, 2006. 2

[22] Fabian Leinen, Vittorio Cozzolino, and Torsten Schön. Volnet: estimating human body part volumes from a single rgb image. *arXiv preprint arXiv:2107.02259*, 2021. 1, 2

[23] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. *Advances in neural information processing systems*, 23, 2010. 2

[24] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018. 2

[25] 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. 1, 3, 4, 5

[26] Dingkang Liang, Wei Xu, and Xiang Bai. An end-to-end transformer model for crowd localization. In *European Conference on Computer Vision*, pages 38–54. Springer, 2022. 2

[27] 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. 1, 2, 3, 4, 5

[28] Chang Liu, Yujie Zhong, Andrew Zisserman, and Weidi Xie. Countr: Transformer-based generalised visual counting. *arXiv preprint arXiv:2208.13721*, 2022. 2

[29] Xingguo Liu, Jianwei Niu, Linghua Ran, and Taijie Liu. Estimation of human body volume (bv) from anthropometric measurements based on three-dimensional (3d) scan technique. *Aesthetic Plastic Surgery*, 41:971–978, 2017. 1, 2

[30] Frank P-W Lo, Yingnan Sun, Jianing Qiu, and Benny Lo. Food volume estimation based on deep learning view synthesis from a single depth map. *Nutrients*, 10(12):2005, 2018. 2

[31] Frank P-W Lo, Yingnan Sun, Jianing Qiu, and Benny PL Lo. Point2volume: A vision-based dietary assessment approach using view synthesis. *IEEE Transactions on Industrial Informatics*, 16(1):577–586, 2019. 2

[32] 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

[33] 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. 1, 3, 4, 5

[34] Zhiheng Ma, Xing Wei, Xiaopeng Hong, Hui Lin, Yunfeng Qiu, and Yihong Gong. Learning to count via unbalanced optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2319–2327, 2021. 2

[35] Aparecido Nilceu Marana, SA Velastin, LF Costa, and RA Lotufo. Estimation of crowd density using image processing. *IET Conference Proceedings*, 1997. 2

[36] World Health Organization et al. Mean bmi (kg/m^2)(crude estimate).[internet]. who. 2017, 2017. 8

[37] 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, 6, 8

[38] Christian Pfitzner, Stefan May, Christian Merkl, Lorenz Breuer, Martin Köhrmann, Joel Braun, Franz Dirauf, and Andreas Nüchter. Libra3d: Body weight estimation for emergency patients in clinical environments with a 3d structured light sensor. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2888–2893, 2015. 1, 2, 7

[39] Katrin Pirker, Matthias Rüther, Horst Bischof, Falko Skrabal, and Georg Pichler. Human body volume estimation in a clinical environment. *AAPR/OAGM: challenges in the biosciences: image analysis and pattern recognition aspects, Stainz Austria*, 2009. 1, 2

[40] Manika Puri, Zhiwei Zhu, Qian Yu, Ajay Divakaran, and Harpreet Sawhney. Recognition and volume estimation of food intake using a mobile device. In *2009 Workshop on Applications of Computer Vision (WACV)*, pages 1–8, 2009. 2

[41] Yasiru Ranasinghe, Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara, and Vishal M Patel. Crowddiff: Multi-hypothesis crowd density estimation using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12809–12819, 2024. 2

[42] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning To Count Everything. pages 3394–3403. 2

[43] GTA Redux. Gta - redux. gta5redux.com, 2023. Accessed on 17 October 2023. 6

[44] Max Roser, Cameron Appel, and Hannah Ritchie. Human height. *Our World in Data*, 2021. https://ourworldindata.org/human-height. 2, 6, 8

[45] 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. 5

[46] Mark P Silverman. Exact statistical distribution of the body mass index (bmi): Analysis and experimental confirmation. *Open Journal of Statistics*, 12(3), 2022. 2, 6, 7

[47] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1221–1231, 2019. 2

[48] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2594–2609, 2020. 2

[49] 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. 1, 2, 3, 4, 5

[50] D Sudharson, J Srinithi, S Akshara, K Abhirami, P Sriharshitha, and K Priyanka. Proactive headcount and suspicious activity detection using yolov8. *Procedia Computer Science*, 230:61–69, 2023. 1

[51] Pongpisit Thanasutives, Ken-ichi Fukui, Masayuki Numao, and Boonserm Kijsirikul. Encoder-Decoder Based Convolutional Neural Networks with Multi-Scale-Aware Modules for Crowd Counting. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2382–2389. 2

[52] Peter A Thompson. Developing new techniques for modelling crowd movement. *KB thesis scanning project 2015*, 1994. 1

[53] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 5, 6

[54] Jia Wan, Ziquan Liu, and Antoni B Chan. A generalized loss function for crowd counting and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1974–1983, 2021. 2

[55] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution matching for crowd counting. *Advances in neural information processing systems*, 33:1595–1607, 2020. 2

[56] 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. 4

[57] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2141–2149, 2020. 2, 3

[58] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[59] 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. 1, 3, 4, 5

[60] Hao Wen, Jing Huang, Huili Cui, Haozhe Lin, Yu-Kun Lai, Lu Fang, and Kun Li. Crowd3d: Towards hundreds of people reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8937–8946, 2023. 3

[61] Chang Xu, Ye He, Nitin Khanna, Carol J Boushey, and Edward J Delp. Model-based food volume estimation using 3d pose. In *2013 IEEE International Conference on Image Processing*, pages 2534–2538. IEEE, 2013. 2

[62] Zhengeng Yang, Hongshan Yu, Shunxin Cao, Qi Xu, Ding Yuan, Hong Zhang, Wenyan Jia, Zhi-Hong Mao, and Mingui Sun. Human-mimetic estimation of food volume from a single-view rgb image using an ai system. *Electronics*, 10(13):1556, 2021. 2

[63] Zhiyuan You, Kai Yang, Wenhan Luo, Xin Lu, Lei Cui, and Xinyi Le. Few-shot Object Counting with Similarity-Aware Feature Enhancement. 2

[64] Shanghang Zhang, Guanhang Wu, Joao P Costeira, and José MF Moura. Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras. In *Proceedings of the IEEE international conference on computer vision*, pages 3667–3676, 2017. 2

[65] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016. 2, 3

[66] Yixuan Zhu, Ao Li, Yansong Tang, Wenliang Zhao, Jie Zhou, and Jiwen Lu. Dpmesh: Exploiting diffusion prior for occluded human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1101–1110, 2024. 3