SAT-HMR: Real-Time Multi-Person 3D Mesh Estimation via Scale-Adaptive Tokens

Supplementary Material

In Sec. A.1, we elaborate on the implementation details of our proposed method and the experimental setups. We provide additional results and extended discussions in Sec. B.

A. Implementation Details

A.1. Model Architecture

Encoder-decoder design. Our model follows an encoder-decoder design based on previous works [2, 23], with the vanilla encoder replaced by our proposed scale-adaptive encoder. For the decoder, following [23], human queries consist of two learnable components: the *content* part and the *positional* part, also known as *decoder embeddings* and *anchor boxes* [23]. Following [23], anchor boxes are refined layer-by-layer by predicting residual values via a prediction head. Additionally, we initialize mean SMPL pose and shape parameters used in [2, 11, 16] and update them with a similar procedure, as illustrated in Fig. A1. Predictions of 3D translation $\boldsymbol{t}$ are regressed from updated decoder embeddings without iteratively updating, which is not included in Fig. A1. Finally, all human predictions are matched to GTs before computing training losses. We adopt the Hungarian algorithm following previous works [5, 44]. The matching cost is computed as a weighted sum of $\mathcal{L}_{\text{box}}$, $\mathcal{L}_{\text{det}}$, and $\mathcal{L}_{\text{j2d}}$, with the weights sharing the same values as those in Sec. A.2.

Camera model. To leverage 2D annotations for supervision, we adopt a pinhole camera model to project 3D joints onto the image plane. Given the focal length f and principal point $(p_{\text{u}}, p_{\text{v}})$, a 3D point (x, y, z) is projected to the image coordinates (u, v) as follows:

$$u = \frac{f \times x}{z} + p_{\text{u}}, \quad v = \frac{f \times y}{z} + p_{\text{v}}. \tag{A1}$$

Following [2, 46], we assume a standard camera with a fixed field of view (FOV) of $60°$. Given S_{hr} as the longer side of the image, the focal length is predefined as $f = S_{\text{hr}}/(2\tan(FOV/2))$. The principal point $(p_{\text{u}}, p_{\text{v}})$ is located at the center of the image.

A.2. Training

The confidence and scale thresholds corresponding to the scale map are set to $\alpha_{\text{c}} = 0.3$ and $\alpha_{\text{s}} = 0.5$, respectively. The loss weights are set to $\lambda_{\text{map}} = 4$, $\lambda_{\text{depth}} = 0.5$, $\lambda_{\text{pose}} = 5$, $\lambda_{\text{shape}} = 3$, $\lambda_{\text{j3d}} = 8$, $\lambda_{\text{j2d}} = 40$, $\lambda_{\text{box}} = 2$ and $\lambda_{\text{det}} = 4$. We train our model with AdamW [25], with weight decay set to $1e-4$. The initial learning rate for the pretrained parameters is set to $2e-5$, while for other parameters, it is set to $4e-5$. The model is trained for 60 epochs with a total batch size of 40, which takes around a week on 8 RTX 3090 GPUs.

A.3. Datasets

We briefly introduce the datasets used for training or evaluation.

AGORA [33] is a synthetic dataset known for its high realism and diverse scenarios. Due to its highly accurate GTs annotated in both SMPL [24] and SMPL-X [34], AGORA has become an essential benchmark for evaluating 3D human mesh estimation models. It contains approximately 14K images with 107K instances for training, 1K images with 8K instances for validation, and 3K images for testing.

BEDLAM [3] is a large-scale, synthetic video dataset that includes a diversity of body shapes, motions, skin tones, hair, and clothing. The dataset contains approximately 286K images with 951K instances for training and 29K images with 96K instances for validation. For our ablation study, we uniformly downsample the training set by a factor of 6, resulting in 48K images with 159K instances. We do not use the test set because the SMPL format is not currently supported by the leaderboard.

COCO [20], **Crowdpose** [17], and **MPII** [1] are real-world multi-person datasets widely used for 2D human pose estimation tasks. We use these datasets for training to enhance the generalization capability of our model on real-world images by using pseudo annotations from NeuralAnnot [31] and only supervise projected 2D joints due to 3D ambiguity and their label noisiness. We uniformly downsample COCO by a factor of 4, resulting in 16K images with 66K instances for training. For Crowdpose, we use 10K images with 36K instances, and for MPII, we use 17K images with 29K instances.

H3.6M [13] is an indoor single-person dataset with 3D pose annotations. It contains videos of common activities performed by professional actors. We uniformly downsample its training set by a factor of 10 and use 31K images.

3DPW [49] is an in-the-wild dataset with 3D mesh annotations. It contains approximately 17K images for training and 24K images for testing. Following [2, 45, 46], we use the training set to finetune our model before evaluating the test set.

MuPoTS [29] is a real-world multi-person 3D pose dataset

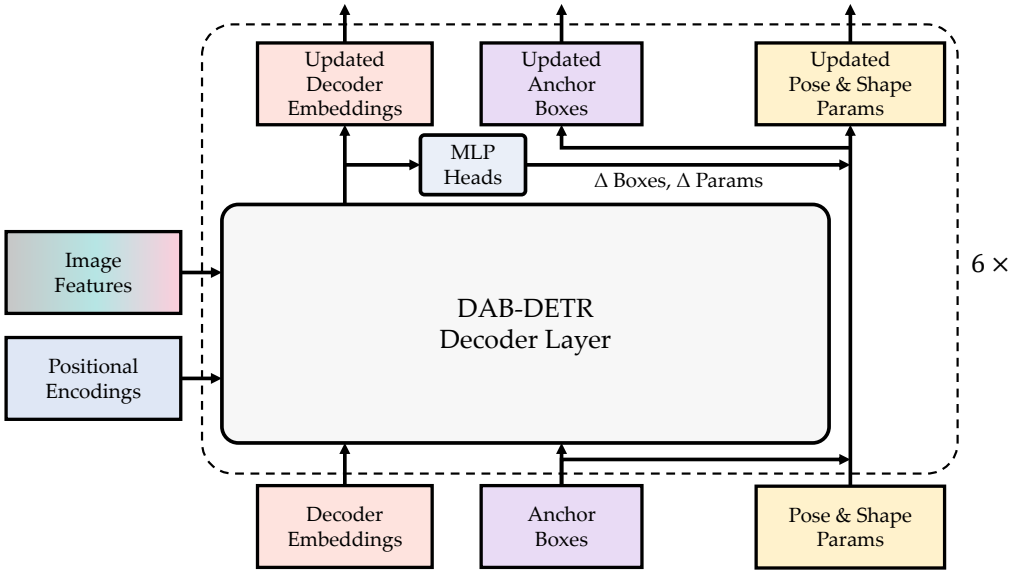

Figure A1. **Illustration of our decoder architecture.** Queries consist of *decoder embeddings* and *anchor boxes* following DAB-DETR [23]. Besides updating anchor boxes, we also update SMPL parameters using corresponding prediction heads.

Table A1. **Speed-accuracy trade-off among different ablation settings.** We conduct our studies on BEDLAM [3] validation set, reporting the average number of different tokens, inference runtime and MVE.

Ablation	Number of tokens			Runtime (ms)	MVE ↓
	High-resolution	Low-resolution	Background		
Single Resolution					
(a) Res. 1288		4784		174.9	53.2
(b) Res. 644		1196		42.3	63.3
Background Tokens $\mathcal{T}_B$					
(c) Drop all	493	40	0	39.7	57.2
(d) No pooling	493	1073		54.1	56.1
(e) Pooling×2	493	94	100	41.6	56.3
(f) **Ours**	493	94	245	42.0	56.0

composed of more than 8K frames from 20 scenes, each containing up to three subjects, annotated with 3D pose. Following previous works [2, 46], we only use it to evaluate the generalization capability of our model.

CMU Panoptic [15] is an indoor multi-person dataset providing 3D pose annotations. It contains 4 sequences of multiple people engaging in different social activities, with approximately 9K images. Following previous works [2, 45, 46], we only use it to evaluate the generalization capability of our model.

B. Extended Results

B.1. Ablation Study

Speed-accuracy trade-off. We evaluate the speed-accuracy trade-off across various ablation models. Specifically, we compare single-resolution baselines with our model, which adopts different processing strategies for background tokens ($\mathcal{T}_B$). In Tab. A1, we report the average number of tokens, inference runtime, and the MVE metric on BEDLAM [3] validation set. The average number of tokens is included to highlight the impact of image tokens on inference speed.

In ablation (a), our baseline model with a resolution of 1288 achieves the lowest estimation error but suffers from redundant image tokens, leading to extremely slow inference. In contrast, ablation (b) shows faster inference but with poor performance, *i.e.* much higher MVE. In (d), simply replacing small-scale tokens with their high-resolution counterparts brings a noticeable boost in performance with additional overhead, where some low-resolution tokens are still redundant. In (f), our proposed method of pooling background tokens $\mathcal{T}_B$ once counteracts the overhead brought by high-

Table B2. **Effect of resolution.** We study the impact of resolution using both single-resolution (baseline) and mixed-resolution (our method, with "*") settings on AGORA [33] and BEDLAM [3], reporting MVE for different scale ranges and the average (Avg.).

	Res.	0-30%	30-50%	50-70%	70%+	Avg.
BEDLAM	644	65.1	61.0	58.4	64.6	63.3
	896	57.8	54.7	54.7	59.3	56.5
	1288	55.2	50.4	51.6	56.0	53.2
	448*	60.3	60.3	61.0	68.6	60.5
	644*	55.6	56.0	57.6	63.1	56.0
	896*	51.1	51.3	53.7	58.6	51.4
	Res.	0-10%	10-20%	20-30%	30%+	Avg.
AGORA	644	100.8	77.2	59.2	53.0	72.0
	896	91.4	71.3	55.3	50.1	67.0
	1288	82.2	64.9	52.4	48.2	61.9
	448*	94.6	74.1	58.5	54.7	70.0
	644*	84.6	68.5	57.3	52.7	65.5
	896*	76.5	63.7	54.1	49.9	61.0

Table B3. **Dataset comparisons.** We report the number of instances for different scale ranges on different datasets, as well as MVE comparisons between our method (with "*") and the baseline.

		0-30%	30-50%	50-70%	70%+	Avg.
BEDLAM	Count	54041	36362	4586	1497	-
	644*	55.6	56.0	57.6	63.1	56.0
	644	65.1	61.0	58.4	64.6	63.3
3DPW	Count	5121	10113	12299	7982	-
	644*	86.2	74.2	70.5	70.2	73.7
	644	93.6	77.8	68.7	68.3	74.8
		0-10%	10-20%	20-30%	30%+	Avg.
AGORA	Count	1556	2976	1986	1274	-
	644*	84.6	68.5	57.3	52.7	65.5
	644	100.8	77.2	59.2	53.0	72.0
		0-40%		40%+		Avg.
Panoptic	Count	31592		11172		-
	644*	85.2		81.0		84.2
	644	90.0		82.6		88.2

resolution tokens, yielding similar performance. However, further reducing $\mathcal{T}_B$ brings no significant acceleration and may result in a potential performance drop, as illustrated in (c) and (e). These results demonstrate that our scale-adaptive strategy achieves the best speed-accuracy trade-off, making our method the **best real-time model**.

Effect of resolution. To study the impact of resolution, we train our single-resolution baseline and our proposed method with scale-adaptive tokens using different resolution settings, reporting estimation errors (MVE) on AGORA [33] validation set and BEDLAM [3] validation set in Tab. B2. As input resolution increases, both single-resolution (baseline) and mixed-resolution (our method, with "*") settings show accuracy improvements across individuals in different scale ranges. Compared to the corresponding baseline, our method greatly reduces the estimation error in small-scale instances. These results further demonstrate the importance of higher resolution, as it leads to better outcomes, and highlight the effectiveness of our mixed-resolution strategy.

Table B4. **Comparison of different scale thresholds on BEDLAM [3] validation set.** We report MVE for different scale ranges, average (Avg.) MVE and inference runtime (ms).

α_s	MVE ↓					Time (ms)
	0-30%	30-50%	50-70%	70%+	Avg.	
(a) 0.0	67.0	60.4	58.6	64.0	64.0	37.2
(b) 0.3	60.0	61.9	59.9	65.8	60.8	40.9
(c) **0.5 (Ours)**	55.6	56.0	57.6	63.1	56.0	42.0
(d) 0.7	55.9	56.1	58.2	65.2	56.2	44.1
(e) 1.0	58.5	56.8	58.5	63.9	57.9	46.4

Scale distribution. We report the number of instances for each scale range on AGORA [33] validation set, BEDLAM [3] validation set, 3DPW [49] test set and CMU Panoptic [15] test set in Tab. B3 (MuPoTS [29] is not included due to the lack of bounding box annotations). To study the impact of scale-adaptive tokens, we evaluate our method (644*, mixed resolution) and the corresponding single-resolution baseline (644) on these datasets. As shown, AGORA [33] and BEDLAM [3] contain more small-scale instances, where our method outperforms baseline with reduced MVE. This improvement is also seen on small-scale instances in 3DPW [49] and CMU Panoptic [15]. However, 3DPW's larger-scale instances show no improvement, likely due to the model's limited capability.

Scale threshold α_s. To further study the impact of high-resolution tokens, we conduct experiments on various scale thresholds α_s while retaining the pooling of background tokens. $\alpha_s = 0$ denotes that no high-resolution tokens are used. Results are shown in Tab. B4. Compared to (a), our method (c) achieves a consistent error reduction across different scale ranges, indicating that introducing sufficient high-resolution tokens eases the estimation challenge on small-scale instances and also allows the model to deal better with large-scale instances. In (b), although the improvement in the scale range of 0-30% is significant, the decrease in high-resolution samples increases learning difficulty during training and potentially leads to worse performance on larger-scale instances than (a). (d) and (e) indicate that too large α_s decreases efficiency with longer interence time cost without

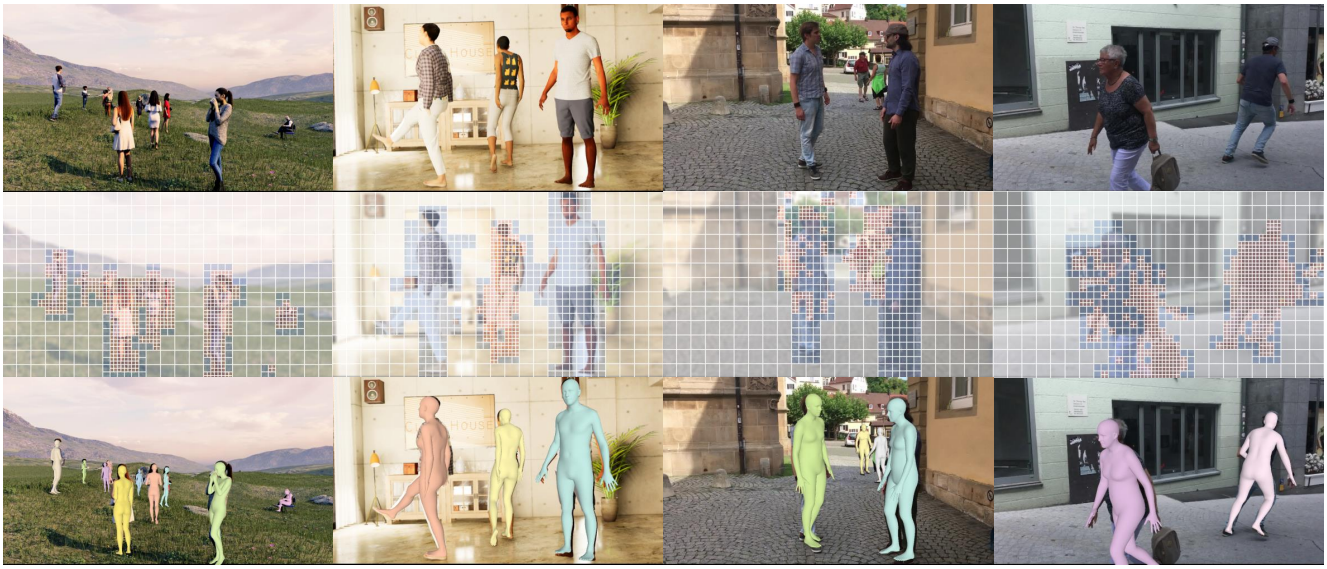

Figure B2. **Additional visualization of scale-adaptive tokens $\mathcal{T}_{\text{SA}}$.** We display the input, scale-adaptive tokens and the estimated mesh overlay (top to bottom). Tokens are visualized the same way as Fig. 6. The first two columns are qualitative results on synthetic datasets [3, 33], and the last two columns are on 3DPW [49].

Table B5. **Accuracy and effect of scale prediction on BEDLAM [3] validation set.** We report F1-Score (F1) and Mean Absolute Error (MAE) for evaluating scale map accuracy. To analyze its impact on mesh estimation, we replace the predicted scale map with GT and report MVE for different scale ranges and the average (Avg.).

Scale map	F1 ↑	MAE ↓	MVE ↓				
			0-30%	30-50%	50-70%	70%+	Avg.
Pred.	0.98	0.056	55.6	56.0	57.6	63.1	56.0
GT	-	-	55.4	55.8	58.6	63.2	55.8

improving accuracy. In (d), a large scale threshold ignores plenty of background context since the high-resolution tokens are encoded independently for N_{hr} layers, leading to a performance decline, which is consistent with our findings in Sec. 4.4. In (e), when all humans are processed with high-resolution tokens (*i.e.* $\alpha_{\text{s}} = 1.0$), training becomes unstable and results are worse. In general, $\alpha_{\text{s}} = 0.5$ achieves the best speed-accuracy trade-off.

Accuracy and effect of scale map prediction. We evaluate our scale map prediction in Tab. B5 which achieves high prediction accuracy, with 0.98 F1-Score. To further analyze its impact on the final mesh estimation, we replace the predicted scale map with GT scale map during inference time. The average MVE slightly improves, indicating our scale predictions are accurate with minimal impact on accuracy. For scales over 50%, we find that the predicted scale map outperforms GT due to underestimating scales in some cases, assigning more instances of high-resolution tokens, thus improving results.

B.2. Additional Qualitative Results

Scale-adaptive tokens. We present additional visualized examples of our scale-adaptive tokens $\mathcal{T}_{\text{SA}}$ in Fig. B2. The last two cases include instances with scales near the scale threshold α_{s}, resulting in mixed-resolution token representations for those individuals. Nevertheless, our model still produces satisfactory predictions, demonstrating the robustness and consistency of the features learned across different resolution levels.

SOTA comparisons. Fig. B3 and Fig. B4 present visual comparisons between our method and existing SOTA approaches [2, 44–46] on synthetic images and real-world images, respectively. Our method demonstrates a strong generalization capability with accurate estimations across different scenarios. Specifically, our method can accurately estimate individuals across different scales, whereas other methods may fail to detect very small individuals or produce inaccurate estimations. See Fig. B4 for qualitative examples illustrating this advantage.

Failure cases. Fig. B5 (top) indicates that our method can result in unsatisfactory depth reasoning without explicit height or age awareness. Fig. B5 (bottom) shows poor mesh estimations on challenging scenes with heavy occlusion and complex human poses, which also challenges existing SOTA methods [2, 44].

B.3. Discussion

Multi-person 3D human mesh estimation is a fundamental task with broad applications. With recent one-stage SOTA methods [2, 44] achieving remarkable improvements in accu-

Figure B3. **Comparison on synthetic images.** We compare our method with other SOTA methods [2, 44] on BEDLAM [3] (left) and AGORA [33] (right). Red dashed circles highlight areas with 2D misalignment or misdetection. The last row shows the elevated view of our estimations. Please zoom in for details.

racy, we further explore the potential of DETR-style pipeline by leveraging scale-adaptive tokens to encode features more efficiently. Our approach achieves superior performance with significantly lower computational cost, marking a step forward for real-time applications. With more diverse training data of high quality GTs, we may further enhance our model's robustness and generalization capability. Additionally, our scale-adaptive tokens may be able to be plugged into other DETR-style works to improve their efficiency in the future.

Limitations. Since our method is not age- or height-aware, it may produce larger depth estimation errors for children, as shown in Fig. B5 (top). In the future, this issue could be addressed by incorporating a mechanism to identify and account for children. Also, we currently only support body-only estimation. Since regions of human face and hands are also challenging and require a higher resolution, our method can be extended to full-body estimation in the future.

Figure B4. **Comparison on real-world images.** We compare our method with SOTA methods [2, 44–46] on in-the-wild images from the Internet. Our method outperforms all of them, especially in small-scale cases. Please zoom in for details.

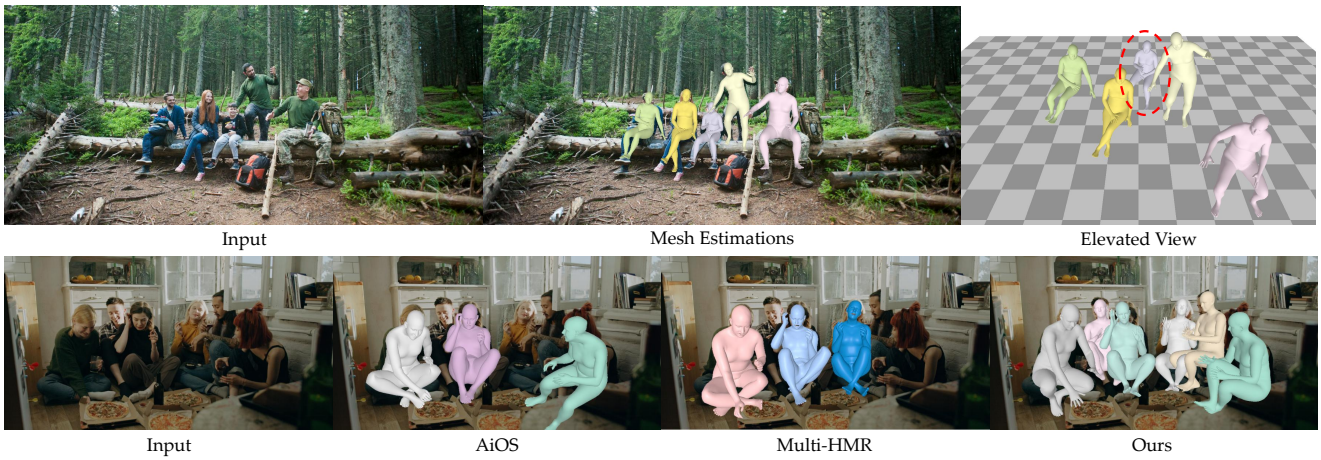

Figure B5. **Failure cases.** The top row shows an example of improper depth reasoning for the child. The bottom row shows poor estimation results of current SOTA methods in complex human poses and scenarios.