WildAvatar: Learning In-the-wild 3D Avatars from the Web

Supplementary Material

A. Details in Data Collection

Downloading Video Candidates. Our first goal is to collect videos from the web with human motions. To cover a wide range of in-the-wild human-central activities, we start from a label pool of human motion datasets [8]. Based on these human motion labels (*e.g.*, fishing and playing tennis), we download over $100k+$ video candidates from YouTube API.

Per-filtering Video Clips. Some collected video candidates could not meet the high-quality human avatar creation requirement. For example, human bodies may not exist at all (*e.g.*, blank preview, severe occlusion) or frequently change across scenes (*e.g.*, montage) in some subsections of these videos. To exclude such unqualified subsections, we utilize SceneDetect [1] to cut video candidates into clips and eliminate those with insufficient length (less than 2 seconds). Subsequently, we apply human detection models with low FPS to these clips to efficiently filter out those without human subjects at minimal cost. After filtering video candidates, we obtain $460k+$ video clip candidates for further processing.

B. Details in Data Pipeline

In addition to the main paper, we provide more details on the data processing pipeline and filtering protocols.

Stage I: Human Bounding Box Detection and Tracking. We first obtain the bounding box of human subjects with off-the-shelf state-of-the-art detection methods (*e.g.*, Yolo [10] and Detectron2 [14]). We only keep the video clip with at least one "person" instance with its detection threshold over 0.8 on all detection models. The tracking step is finding the largest IOU overlay of bounding boxes among frames. We discard low-resolution human subjects whose bounding box areas are lower than 64×64. To ensure the richness of the dataset, we only keep one "key subject" for each video, as clips from the same video may probably share the same key subject.

Stage II: Human Segmentation Mask Extraction. We first obtain the 2D keypoints J_{2D} for human subjects using the popular HRNet [12] and DWPose [16]. Given the 2D keypoint annotations, we can also discard over part-occluded subjects. In particular, we only keep the subject with the average confidence of 2D keypoints over 0.65. For segmentation, we feed the 2D bounding box and the 2D keypoints into the *sam_vit_h* sub-model to extract the foreground mask.

Stage III: Coarse SMPL and Camera Estimation. We first estimate SMPL and camera parameters frame by frame, using state-of-the-art single-image-based human pose and shape (HPS) estimation methods [3, 7]. To perform better in complex scenes in the wild, we adapt the model pre-trained on the in-the-wild 3DPW dataset [13]. The HPS models infer human body pose/shape parameters (θ/β) and the global camera parameters (rotation matrix R and the 3D offset T). To retain the remaining video clips with considerable viewpoint shifts and human movements, we discard the clips with viewpoint angle changes lower than $\frac{\pi}{4}\ rad$. We also automatically select the most non-trivial $N = 20$ frames, which keeps the pose and viewpoint diversity to the greatest extent possible. As mentioned in the main paper, we double-check the consistency of the SAM and SMPL masks. Intuitively, the SMPL mask denotes the naked body, while the SAM mask contains the clothed body. Therefore, the SMPL mask should be mostly covered by the SAM mask (See Fig. F (a) ~ (d)). We discard the subjects whose SAM masks are over $3\times$ larger than their SMPL masks (See Fig. F (e) ~ (h)). We also discard the subjects whose 10% SMPL mask pixels from main bodies are not covered by the SAM mask (See Fig. F (i) ~ (l)). Similarly, we double-check the consistency of the 2D keypoints from 2D pose and SMPL estimations and discard the clips with the averaged PCK less than 0.85.

Stage IV: Refining SMPL and Camera In-the-loop. We refine the coarse SMPL parameters (θ, β) and camera parameters (R, T) obtained in Stage I for high-quality annotations. To achieve temporally smooth results, we regularize the differences in parameters between adjacent frames, which is given by

$$
\begin{aligned}
\mathcal{L}^s_\theta &= \sum_{i=1}^{N-1} \left\| \theta^i - \theta^{i+1} \right\|_2, \\
\mathcal{L}^s_R &= \sum_{i=1}^{N-1} \left\| R^i - R^{i+1} \right\|_2, \\
\mathcal{L}^s_T &= \sum_{i=1}^{N-1} \left\| T^i - T^{i+1} \right\|_2, \\
\mathcal{L}^s_{2D} &= \sum_{i=1}^{N-1} \Big\| \Pi(J_{3D}(\theta^i, \beta^i); R^i, T^i) \\
&\qquad - \Pi(J_{3D}(\theta^{i+1}, \beta^{i+1}); R^{i+1}, T^{i+1}) \Big\|_2,
\end{aligned}
\tag{1}
$$

where $J_{3D}(\theta, \beta)$ infers the 3D keypoints of the human body, and the Π denotes the 2D projection, and i denotes the i_{th} frame of the input video. Notice that the body shapes (β) are treated as constants across the input video.

In addition, we align the human body parameters to the

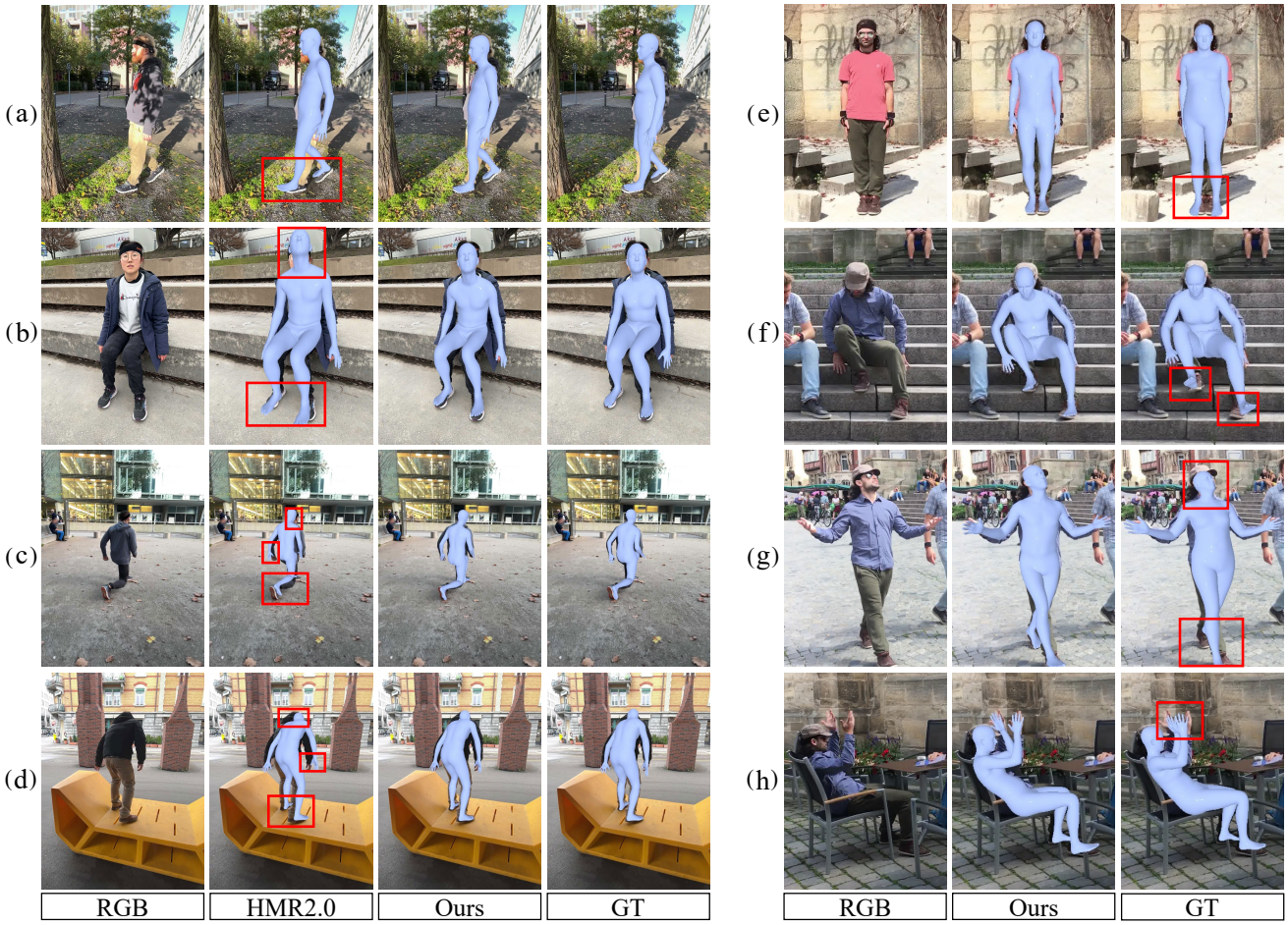

Figure A. Qualitative comparison of our four-stage pipeline and the state-of-the-art HMR2.0 [3]. Our pipeline can adapt to complex environmental scenes and output reasonable results in uncommon scenarios.

2D keypoints and SAM masks, which are given by

$$\mathcal{L}_{2D} = \sum_{i=1}^{N} \left\| \Pi(J_{3D}(\theta^i, \beta^i)) - J_{2D}^i \right\|_2, \\ \mathcal{L}_{\text{mask}} = \sum_{i=1}^{N} (M_{rd}(\theta^i, \beta^i)), M^i), \tag{2}$$

where $M_{rd}(\theta, \beta)$ denotes the rendered mask related to the SMPL parameters. The M^i denotes the foreground of the i frame. We also regularize θ to avoid out-of-domain poses [2], using the Gaussian Mixture Model (GMM) prior [9]

$$\mathcal{L}_{\text{prior}} = \sum_{i=1}^{N} \left\| GMM(\theta^i) \right\|_2. \tag{3}$$

We adopt the loss functions as mentioned above for supervision:

$$\mathcal{L} = \lambda_\theta^s \mathcal{L}_\theta^s + \lambda_R^s \mathcal{L}_R^s + \lambda_T^s \mathcal{L}_T^s + \lambda_{2D}^s \mathcal{L}_{2D}^s \\ + \lambda_{2D} \mathcal{L}_{2D} + \lambda_{\text{mask}} \mathcal{L}_{\text{mask}} + \lambda_{\text{prior}} \mathcal{L}_{\text{prior}}. \tag{4}$$

We empirically set the loss weights as $\lambda_\theta^s = 100$, $\lambda_R^s = 1000$, $\lambda_T^s = 50$, $\lambda_{2D}^s = 100$, $\lambda_{2D} = 100$, $\lambda_{\text{mask}} = 100$ and $\lambda_{\text{prior}} = 0.1$. We adopt the LBFGS [11] optimizer with the learning rate $lr = 1.0$. Qualitative comparisons between coarse and refined SMPL and camera parameters can be found in Fig. E.

Efficiency. Data collection and annotation efficiency are also crucial for data scale-up and application. Despite the complex multiple-stage design, our data-collecting pipeline only takes less than 200 seconds to generate full annotations for a 20 seconds in-the-wild video clip.

C. Details in Pipeline Evaluation

Dataset. We evaluate our four-stage pipeline on the in-the-wild EMDB [6], which is widely recognized for its challenging and diverse real-world scenarios. We use the EMDB-1 split to evaluate the camera-coordinate performance. EMDB-1 contains 17 sequences totaling 13.5 minutes.

Metrics. For joints, we compute error on the 24 main joints of the human body under the SMPL convention. As for vertex, we calculate the point-to-point corresponding error

Figure B. Visualization of SMPL overlay on unusual camera viewpoints. Our pipeline can show robust annotations on unusual camera viewpoints and body poses.

on the SMPL vertices. We report quantitative results on MPJPE, PA-MPJPE [5], and PVE [15]. ***MPJPE (Mean Per Joint Position Error)*** calculates the mean distances between the predicted and ground-truth 3D joints after the translation alignment at the pelvis joint. The predicted or ground-truth 3D joints are regressed from corresponding pose and shape parameters. ***PA-MPJPE (Procrustes analysis MPJPE)*** calculates the mean distances between the predicted and ground-truth 3D joints after Procrustes Analysis [4], including alignment in scale, translation and rotation. PA-MPJPE mainly focuses on the quality of pose and shape estimation, regardless of global rotation. ***PVE (Per Vertex Error)*** calculates the mean distances between the vertices on the human mesh without any alignment, which evaluates the reconstruction accuracy of the human surface.

Qualitative Comparisons. Qualitative comparisons on the EMDB dataset are also shown in Fig. A (a) ~ (d). Compared to previous state-of-the-art HMR2.0 [3], our pipeline can adapt to complex environmental scenes and output reasonable results in uncommon scenarios. For example, our pipeline can accurately predict 1) the foot and ankle pose of Fig. A (a) & (c); 2) the head and neck pose of Fig. A (b) & (d); 3) the global alignment of Fig. A (a), (b) & (d). Qualitative comparisons on the 3DPW dataset are also shown in Fig. A (e) ~ (h). Compared to the official ground truth (GT) from expensive IMUs, our annotations also exhibit better alignment on 1) the foot and ankle pose of Fig. A (e) ~ (g); 2) the head and neck pose of Fig. A (g); 3) the hand pose of Fig. A (h). These comparisons validate the annotation quality of our pipeline for in-the-wild videos.

D. License, Statistics and Visualizations

The authors bear all responsibility in case of violation of rights and confirm that this dataset is open-sourced under the **S-Lab License 1.0 license**. We shall enforce strict regulations when applying our code and data to mitigate potential negative social impacts. 69.1% / 30.9% scenes in WildAvatar are indoor/outdoor, respectively. 45.3% / 54.7% of the scenes in WildAvatar have single/multiple human(s), respectively. And 34.6% / 65.4% subjects in WildAvatar are male/female, respectively. More visualization of SMPL overlay on unusual camera viewpoints can be found in Fig. B. More RGB examples can be found in Fig. C, and examples of different SSIOU ranges can be found in Fig. D.

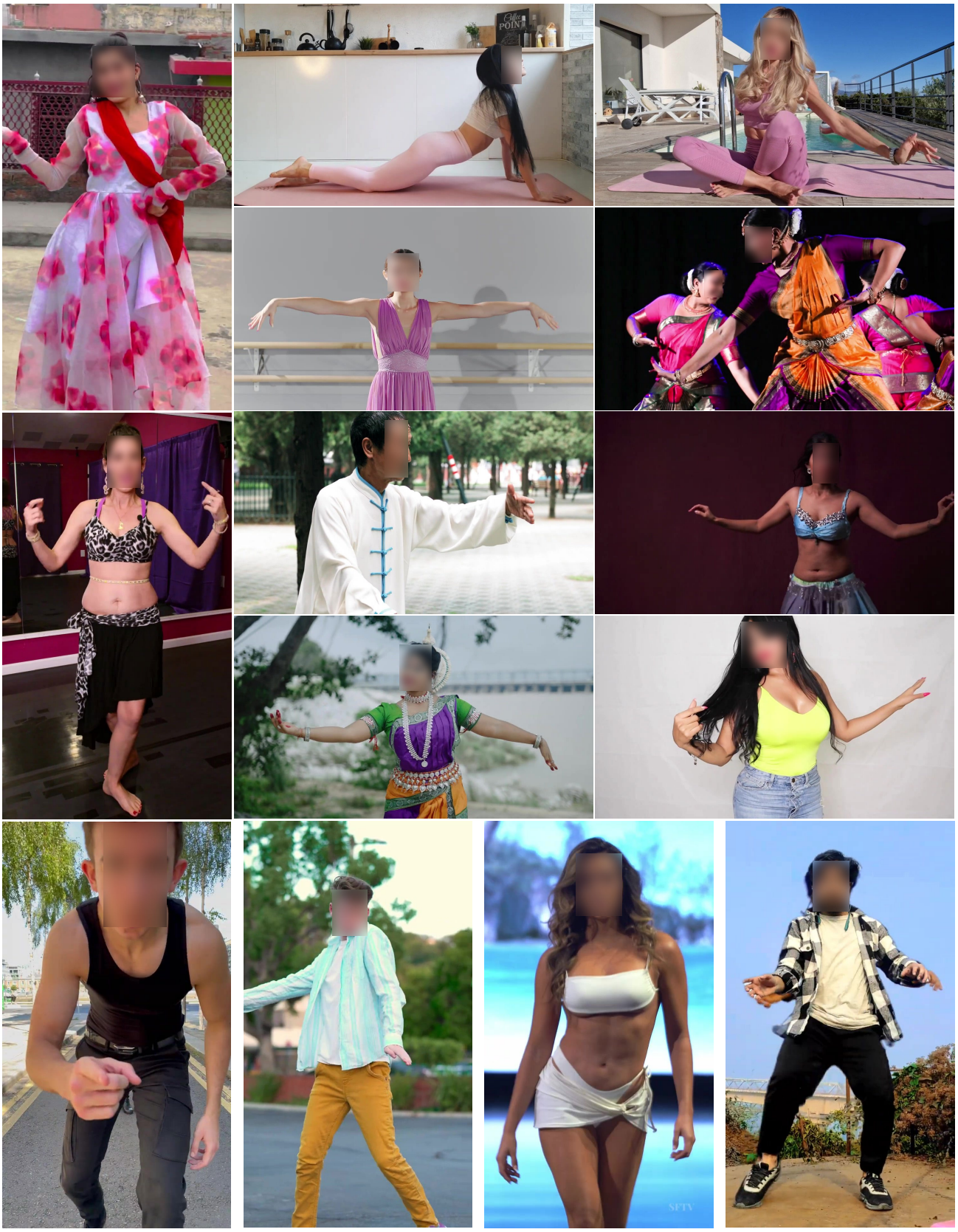

Figure C. More RGB examples from the proposed WildAvatar dataset. The best view zoomed in on-screen for details.

Figure D. Examples of different SSIOU ranges. SSIOU rises from up to down. The last row shows the SSIOU larger than 1.9. The samples with SSIOU larger than 1.9 are mostly caused by loose dresses rather than erroneous SMPL fitting.

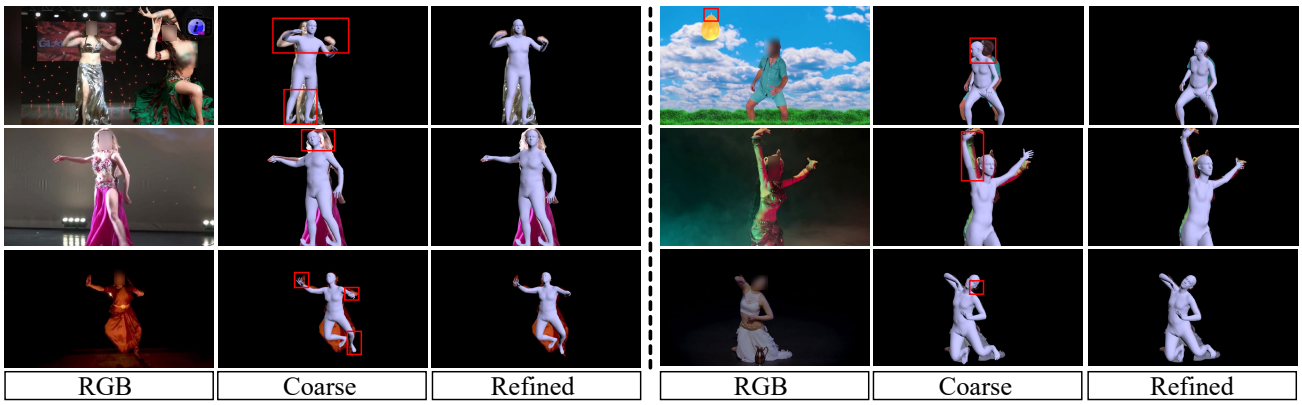

Figure E. Comparison of the coarse and refined SMPL parameters. The coarse SMPL annotations are from Stage III, and are later refined in Stage IV. The refined SMPL parameters achieve better alignments to the raw RGB images.

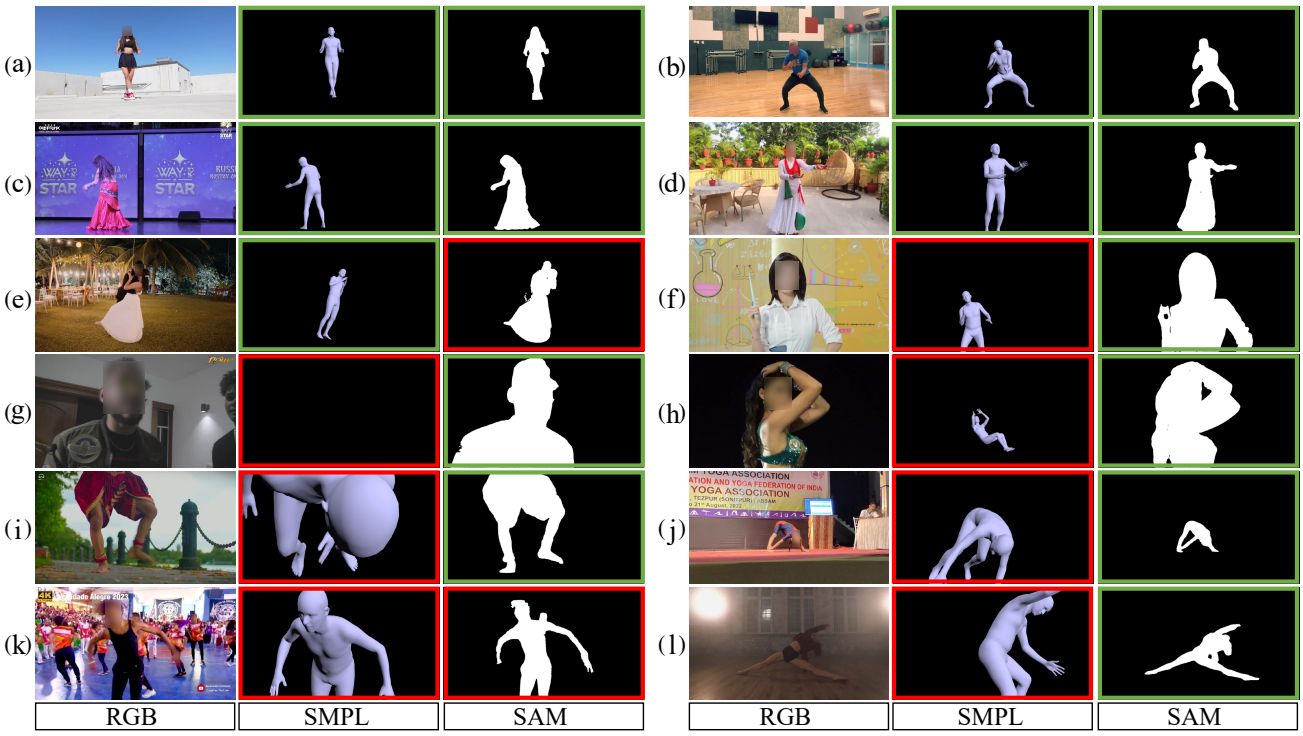

Figure F. SMPL and SAM consistency. The green/red borders denote good/bad outputs, respectively. Note that the SMPL annotations are coarse results from Stage III, which are later refined in Stage IV.

References

[1] Video scene cut detection and analysis tool. Github, 2014. 1

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: automatic estimation of 3d human pose and shape from a single image. In *Proc. Eur. Conf. Comput. Vis.*, pages 561–578. Springer-Verlag, 2016. 2

[3] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 14737–14748. IEEE, 2023. 1, 2, 3

[4] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40:33–51, 1975. 3

[5] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2014. 3

[6] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zarate, and Otmar Hilliges. EMDB: the electromagnetic database of global 3d human pose and shape in the wild. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 14586–14597. IEEE, 2023. 2

[7] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: part attention regressor for 3d human body estimation. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 11107–11117. Computer Vision Foundation / IEEE, 2021. 1

[8] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 1

[9] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proc. IEEE Conf. Comput. Vis. Patt. Recogn.*, pages 10975–10985. Computer Vision Foundation / IEEE, 2019. 2

[10] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 779–788. IEEE Computer Society, 2016. 1

[11] M. Schmidt. minfunc: unconstrained differentiable multivariate optimization in matlab. 2005. 2

[12] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proc. IEEE Conf. Comput. Vis. Patt. Recogn.*, pages 5693–5703. Computer Vision Foundation / IEEE, 2019. 1

[13] 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 *Proc. Eur. Conf. Comput. Vis.*, pages 614–631. Springer-Verlag, 2018. 1

[14] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 1

[15] Youze Xue, Jiansheng Chen, Yudong Zhang, Cheng Yu, Huimin Ma, and Hongbing Ma. 3d human mesh reconstruction by learning to sample joint adaptive tokens for transformers. In *Proc. ACM Int. Conf. Multimedia*, pages 6765–6773. ACM Press, 2022. 3

[16] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pages 4212–4222. IEEE, 2023. 1