Multi-RoI Human Mesh Recovery with Camera Consistency and Contrastive Losses –Supplementary Material–

Yongwei Nie[1], Changzhen Liu[1], Chengjiang Long[2], Qing Zhang[3], Guiqing Li[1], and Hongming Cai[1]

[1] South China University of Technology, China
[2] Meta Reality Labs, USA
[3] Sun Yat-sen University, China
{nieyongwei}@scut.edu.cn

Abstract. In this supplementary material, we first explain the extraction of multiple RoIs in detail and present experiment for different extraction strategies. We also provide details of the RoI-Aware Fusion (RAF) network and the projector network for contrastive learning. In addition, we provide more qualitative results on various datasets.

1 Details of Selecting RoIs

One of our contributions is the multiple RoIs that enable us to utilize the relations between different RoIs of the same human in an image. In this section, we illustrate in detail how we perform the extraction of RoIs and comparison for different extraction strategies.

1.1 Fixed Boundingboxes

The default method is illustrated in Figure 1, where (a) shows a full image containing a target human and its original boundingbox,

$$\mathbf{B}_1 = (c_x, c_y, b). \tag{1}$$

The symbols c_x and c_y denote the distance between the center of the boundingbox and the center of the full image, while the symbol b is the width of the boundingbox. Based on the original boundingbox, we generate other four boundingboxes in a fixed manner, as shown in Figure 1 (b)-(e). The boundingbox in Figure 1 (b) is:

$$\mathbf{B}_2 = (c_x + 0.1b, c_y, 1.5b), \tag{2}$$

i.e., we slightly translate the original boundingbox in the x-y plane by $(0.1b, 0)$, and enlarge the size of the boundingbox by a factor of 1.5. Similarly, all the other boundingboxes are generated by:

$$\begin{aligned} \mathbf{B}_3 &= (c_x - 0.1b, c_y, 1.25b), \\ \mathbf{B}_4 &= (c_x, c_y + 0.1b, 0.8b), \\ \mathbf{B}_5 &= (c_x, c_y - 0.1b, 0.65b) \end{aligned} \tag{3}$$

Together with the original boundingbox, we obtain $M = 5$ boundingboxes. After generating the M boundingboxes, we crop the full image M times using the boundingboxes as RoIs, and input the M RoIs to the proposed network.

1.2 Random Boundingboxes

By default, our method uses fixed boundingboxes. We have also conducted an experiment with random boundingboxes. The random boundingboxes is obtained by:

$$\mathbf{B}_i = (c_x + x_i, c_y + y_i, s_i * b), \tag{4}$$

where $x_i \in [-0.1b, 0.1b]$, $y_i \in [-0.1b, 0.1b]$ and $s_i \in [0.65, 1.5]$ are randomly generated values in the corresponding range. We use random boundingbox for RoIs in both training and testing phases.

1.3 Ablation on Extraction Strategies

In this section, we present the ablation study on different extraction strategies discussed above, *i.e.*, Fixed or Random Boundingboxes. By default, both resizing and translating are used when extracting multiple RoIs in a Fixed manner as in Section 1.1. Hence, we additionally conduct experiments to prove the necessity for both of these two schemes.

Table 1 shows different extraction strategies and the corresponding performance of estimated results using different schemes. To start with, "Fixed" w/o resizing and w/o translating in the first row simply repeats the original boundingbox multiple times. Network trained with repeated boundingboxes obtains worse performance compared to those with only translating ("Fixed w/o resizing") or only resizing ("Fixed w/o translating") due to lack of abundant pairwise relation information between different RoIs. Resizing and translating are both beneficial as they can enrich the visual details around the target person which assist the network to estimate the camera parameters and provide pairwise relation between different RoIs. Random boundingboxes discussed in Section 1.2 obtains worst performance among all the strategies since it may increase the difficulty of network stably learning the pairwise relation. Moreover, it makes the RoIs of test samples not consistent with training ones as the boundingbox are randomized. As seen, the default manner ("Fixed w/ resizing and w/ translating") gives the best result over all strategies.

2 Network Details

2.1 RoI-Aware Fusion Network (RAF)

As shown in Figure 2, we illustrate the detailed architecture of our RAF network. Overall, we concatenate $\{\mathbf{h}_n\}_{n=1}^{M}$ and $\{\gamma_{mn}\}_{n=1}^{M}$ together, and send them to an MLP block to get relation weights $\{w_{mn}\}_{n=1}^{M}$, where $w_{mn} \in [0, 1]$. Specifically, let $M \times d_f$ be the size of the concatenated features, we first downsample the

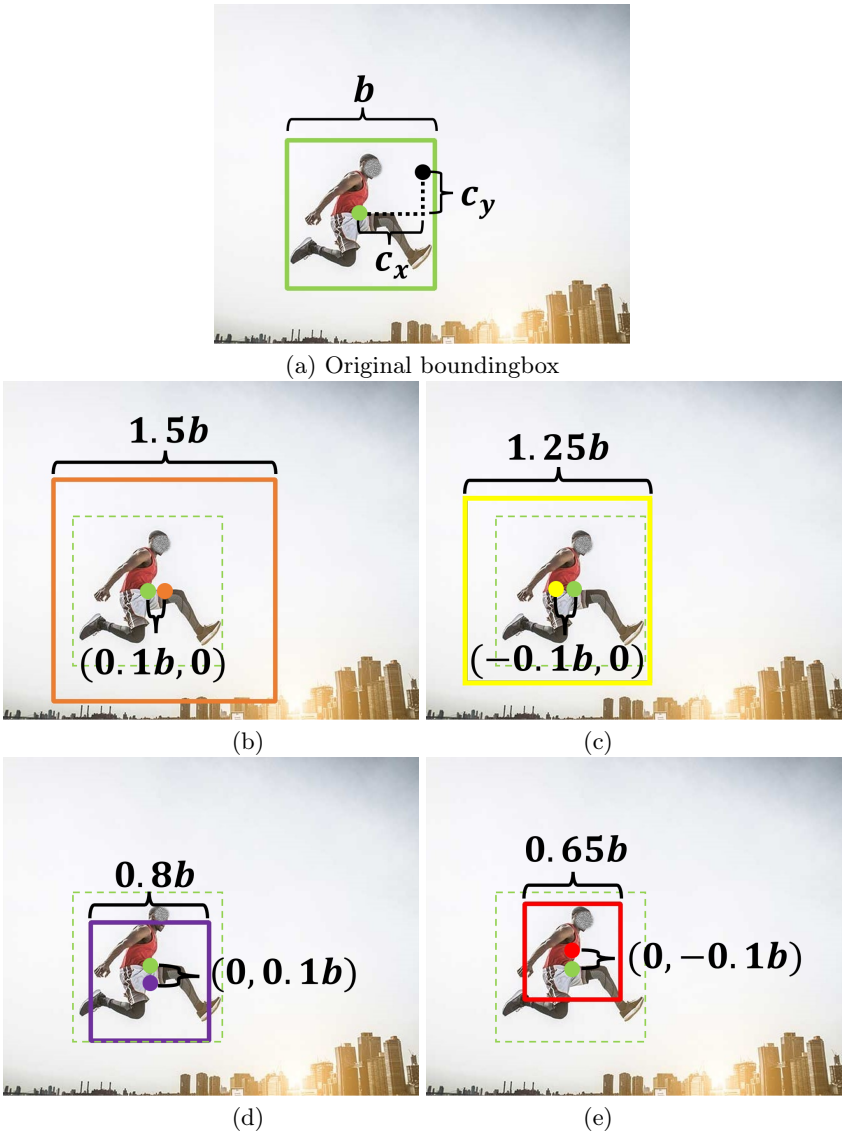

Fig. 1: Boundingbox Definition. (a) Original boundingbox $\mathbf{B}_1 = (c_x, c_y, b)$, where (c_x, c_y) is the distance from the boundingbox center to the image center and b is the width of the boundingbox. (b) Boundingbox $\mathbf{B}_2 = (c_x + 0.1b, c_y, 1.5b)$, which is obtained by moving $\mathbf{B}_1$ along the x-axis by 0.1b and then resizing it by a factor of 1.5. (c) Similarly, $\mathbf{B}_3 = (c_x - 0.1b, c_y, 1.25b)$. (d) $\mathbf{B}_4 = (c_x, c_y + 0.1b, 0.8b)$. (e) $\mathbf{B}_5 = (c_x, c_y - 0.1b, 0.65b)$.

last channel of these features to 256 and flatten them to one feature of $256 \cdot M$ channels. Then we use multiple fully-connected layers, each followed by Tanh

Table 1: Ablation on extraction strategies. "Fixed": the relative positions of boundingbox for RoIs with respect to the originally detected one are the same across different training/testing samples. "Random": the relative positions are different per sample.

Boundingbox	Resizing	Translating	MPJPE	PA-MPJPE
Fixed	✗	✗	83.6	54.4
	✓	✗	83.4	53.3
	✗	✓	83.1	52.0
	✓	✓	**80.8**	**51.9**
Random	✓	✓	84.4	54.1

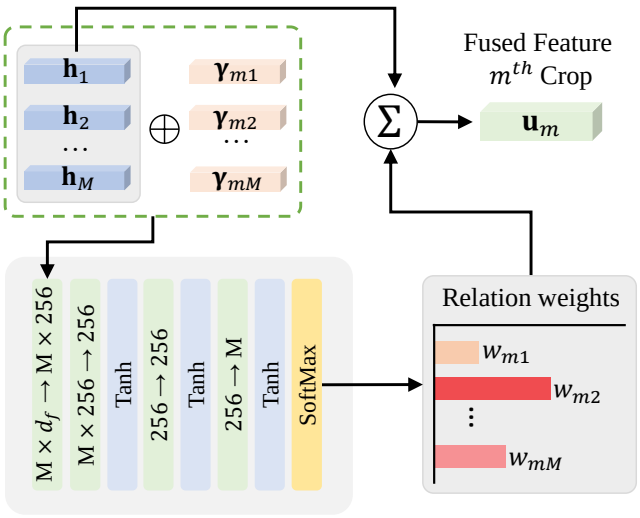

Fig. 2: Detailed architecture of the RAF network.

activation, to gradually reduce the feature size to M. Finally, the M values is normalized by a Softmax, yielding a vector of weight $\{w_{mn}\}_{n=1}^{M}$.

2.2 Projector Network for Contrastive Learning

In our implementation, the contrastive learning model is mainly borrowed from SimCLR [1] with a slight modification by adding a weighting module between the feature extractor E and projector g.

Specifically, a feature vector $\mathbf{h}_m \in \mathbb{R}^d$, $m \in [1, M]$, which is extracted from a cropped image via E, is sent to a fully-connected layer and a Sigmoid activation function, obtaining weights $\mathbf{W}_m \in [0, 1]^d$. We multiply $\mathbf{h}_m$ with $\mathbf{W}_m$, and then send the weighted feature to the projector g to obtain feature $\mathbf{z}_m \in \mathbb{R}^d$.

$$\mathbf{z}_m^i = g(\mathbf{W}_m^i \cdot \mathbf{h}_m^i). \tag{5}$$

Table 2: Complete ablation on components of our method.

No.	$\mathcal{L}_{cont}$	RAF	$\mathcal{L}_{cam}$	MPJPE	PA-MPJPE
(1)	✗	✗	✗	87.0	55.8
(2)	✓	✗	✗	85.9	55.1
(3)	✗	✗	✓	85.2	54.5
(4)	✗	✓	✗	83.0	53.5
(5)	✗	✓	✓	83.4	53.3
(6)	✓	✗	✓	83.8	54.8
(7)	✓	✓	✗	83.1	52.0
(8)	✓	✓	✓	**80.8**	**51.9**

As for the feature projector g, we follow the architecture from [1]. It is composed of 2 blocks with a ReLU activation in between and each block contains a fully-connected layer and a normalization layer. Eventually, we get $\{\mathbf{z}_m\}_{m=1}^{M}$.

3 Ablation on Removing One, Two or Three Design Components

In this section, we demonstrate the complete ablation experiments on our three key components, including RAF, $\mathcal{L}_{cam}$ (camera consistency loss) and $\mathcal{L}_{cont}$ (contrastive loss), as shown in Table 2. In the main paper, we only show experiments of removing one or three components. Here, we additionally show results after removing two components. In the following, we use Row (1)&(2) to represent the comparison between Row (1) and Row (2) in convenience. All of these models are trained on the COCO [5] training dataset and evaluated on 3DPW testing dataset.

Row (1) is the multi-RoI model without all the three components we propose and just uses the averaged $\{\mathbf{h}_m\}_{m=1}^{M}$ instead of $\{\mathbf{u}_m\}_{m=1}^{M}$ to regress the mesh, which can be viewed as a Multi-RoI baseline network. From Row (1)&(4) and Row (6)&(8), we observe that adding RAF network for fusion brings a significant drop of both PA-MPJPE and MPJPE, which indicates the effectiveness of our RAF network. From Row (1)&(3) and Row (7)&(8), we observe that camera consistency contributes more to the drop in MPJPE, as with $\mathcal{L}_{cam}$ we can obtain more accurate global rotation. From Row (3)&(6), we observe that contrastive learning scheme helps the network regress more accurate poses regardless of root rotations and translations, which consequently reduces the PA-MPJPE after alignment. The experiments in Table 2 validate the effectiveness of each of the proposed design components in our method.

4 More Qualitative Results

In Figure 3, we display more qualitative examples on test/validation sets of COCO-EFT [3], 3DPW and Human 3.6M with ground-truth 3D meshes in green.

The cases are challenging because of unusual poses, truncation or body-part occlusion, where our method shows its robustness and produces more reasonable estimation results.

Fig. 3: More examples of qualitative comparison with SOTA approaches. These cases are challenging because of unusual poses, truncation or body-part occlusion.

References

1. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: Simclr: A simple framework for contrastive learning of visual representations. In: International Conference on Learning Representations. vol. 2 (2020)
2. Cho, J., Youwang, K., Oh, T.H.: Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In: European Conference on Computer Vision. pp. 342–359. Springer (2022)
3. Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In: 2021 International Conference on 3D Vision (3DV). pp. 42–52. IEEE (2021)
4. Li, Z., Liu, J., Zhang, Z., Xu, S., Yan, Y.: Cliff: Carrying location information in full frames into human pose and shape estimation. In: European Conference on Computer Vision. pp. 590–606. Springer (2022)
5. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
6. Wang, Y., Daniilidis, K.: Refit: Recurrent fitting network for 3d human recovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14644–14654 (2023)