Pose-Guided Temporal Enhancement for Robust Low-Resolution Hand Reconstruction

Supplementary Material

1. Overview

We first provide more implementation details in Sec. 2, and then introduce the partitioning design of pose-aware tokens in Sec. 3. Furthermore, we have supplemented the qualitative results before Procrustes alignment in Sec. 4. Finally, we present more qualitative results in Sec. 6.

2. Implementation

Network Architecture We have modified ViTPose-B [4] as the backbone of the temporal enhancement framework, setting the dimensions and layers of tokens in ViT to 768 and 12, respectively. We further replaced patch embedding and position encoding in the network with two-layer convolution layers with zero padding [2], as shown in Table 1.

Layer	Kernel	Output channel	Stride	Padding
1	7×7	64	4	2
2	3×3	768	2	1

Table 1. Convolution Layer Setting

Regression Hand Mesh We adopt a regression-based method in [5] to regress vertex-wise 2.5D heatmap, which is represented as a latent 2D heatmaps $B_k \in \mathbb{R}^{h_u \times w_u}$ and a depthmap $D_k \in \mathbb{R}^{h_u \times w_u}$. The 2.5D vertex coordinates are obtained from 2D heatmap using soft-argmax:

$$p_k = \sum_p softmax(B_k(p)) \cdot p \tag{1}$$

$$z_k = \sum_p softmax(B_k(p)) \cdot D_k(p) \tag{2}$$

where p donates the pixel location in heatmap, p_k donates the pixel location in image, z_k donates root-relative depth.

The vertexs' 3D coordinates in camera coordinate can be recovered from above 2D pixel coordinate and relative according to the following relationship:

$$(x_k, y_k, z_k)^T = (z_k^r + z_r)\mathcal{K}^{-1}(u_k, v_k, 1)^T, k = \{0, \ldots, n-1\} \tag{3}$$

where (x_k, y_k, z_k) is the 3D coordinates, (u_k, v_k) is 2D pixel coordinates p_k , z_k is relative depth, z_r is root depth, and $\mathcal{K}$ is camera intrinsic parameters.

Methods	VIBE+LR ViT	**ours-single**	**ours**	**ours(32)**
MPJPE	18.22	14.05	13.21	14.56
MPVPE	15.41	13.54	12.73	14.00

Table 2. Comparison with the state-of-the-art on the DexYCB dataset(before **PA**).

Methods	VIBE + LR ViT	**ours-single**	**ours**	**ours(32)**
MPJPE	13.25	12.50	10.75	11.90
MPVPE	12.50	12.34	10.69	11.76

Table 3. Comparison with the state-of-the-art on the HanCo dataset(before **PA**).

	w/o conv		w/ conv	
Methods	J-PA	V-PA	J-PA	V-PA
Single Plane	6.03	6.07	5.85	5.90
Triplane	5.79	5.84	5.77	5.74

Table 4. Ablation study on convolution layers in Triplane Feature Encoding on the HanCo dataset.

Implementation of Triplane Encoding We use the 2.5D coordinates of the joints to generate distance dependent heatmaps for spreading joint features across different planes. For each joint, we generate a heatmap for each plane. Taking the xy-plane as an example:

$$H_{x,y} = softmax((K - \|p_{x,y} - p_j\|_2)/K) \tag{4}$$

where $p_{x,y}$ is the pixel index of the heatmap, p_j is the projection coordinate of the j_th joint on the xy-plane.

After obtaining the heatmap of each joint in the three planes, taking the xy plane as an example, we obtain the following plane features:

$$F_{x,y}(p) = \sum_j H_{x,y}(p) \cdot J_j \tag{5}$$

where p is the heatmap index, J is the fused joint feature, and j is the joint index.

3. Design of Pose-Aware Tokens

We partition different pose-aware tokens according to the semantic structure of the hand. Specifically, each token independently tracks the palm and individual fingers. Their mapping relationship with specific joints is as follows:

Figure 1. Qualitative Results in the wild. The original images were captured at a resolution of 1920×1080, with the distance between the person and the camera being greater than 2 meters. The cropped hand bounding boxes have an average resolution of less than 64×64.

$$
\begin{aligned}
T_p^0 &\rightarrow \{wrist\} \\
\left.\begin{aligned}
T_p^1 &\rightarrow \{thumb_i\} \\
T_p^2 &\rightarrow \{index_i\} \\
T_p^3 &\rightarrow \{middle_i\} \\
T_p^4 &\rightarrow \{ring_i\} \\
T_p^5 &\rightarrow \{pinky_i\}
\end{aligned}\right\} & \; i \in \{cmc, mcp, ip, tip\}
\end{aligned} \tag{6}
$$

During pose aware fusion, these tokens are split into pose template based on this mapping relationship, which are used to fuse 3D joint features.

4. More Quantitative Results

As results of hand mesh estimation before procrustes alignment are also significantly important in the literature, We further present the evaluation metrics before alignment on DexYCB and HanCo in Tab. 2, Tab. 3. We compared the results with our benchmark VIBE [3] in low resolution. It can be seen that our method can still align images well at low resolutions.

5. Ablation Study

During the process of fusing Triplane features, we applied convolutional layers to capture the local spatial information within each plane. In this section, we investigate the impact of incorporating convolutional layers on the overall performance of the framework. Specifically, we compare the fusion of Triplane features with and without convolutional layers. Additionally, we evaluate the effect of convolutional layers on the projection of single-plane features. The results of these ablation experiments are presented in Tab. 4. The experimental findings demonstrate that the introduction of convolutional layers is effective.

6. Qualitative results

In this PDF, we present an exemplary set of our hand pose estimation sequence on the DexYCB [1] dataset in Fig. 2 and on the HanCo [6] dataset in Fig. 3. The qualitative results shows that that our method is generally robust in low resolution images. We also present the qualitative results on in-the-wild low resolution images in Figure 1.

References

[1] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 2

[2] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021. 1

[3] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 2

[4] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 1

[5] Xiaozheng Zheng, Pengfei Ren, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. Sar: Spatial-aware regression for 3d hand pose and mesh reconstruction from a monocular rgb image. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 99–108. IEEE, 2021. 1

[6] Christian Zimmermann, Max Argus, and Thomas Brox. Contrastive representation learning for hand shape estimation. In *Pattern Recognition*, pages 250–264, Cham, 2021. Springer International Publishing. 2

Figure 2. Qualitative results on the DexYCB dataset.

Figure 3. Qualitative results on the HanCo dataset.