HandGCNFormer: A Novel Topology-Aware Transformer Network for 3D Hand Pose Estimation

Supplementary material

This supplementary material contains the following details: (1) A brief description of our baseline. (2) Additional ablation studies. (3) Additional visualization results.

A. Baseline Architecture

We elaborate on the baseline discussed in Section 4.2 of the main paper. The architecture of our baseline is shown in Figure 1. Following DETR [1], our baseline consists of a ResNet [2] backbone, a Transformer encoder, a Transformer decoder, and a MLP head. Given a 256×256 cropped image, the ResNet and the Transformer encoder are utilized to extract features in the same manner as our HandGCNFormer, resulting in a memory $\mathbf{M}$ that is provided to a standard non-autoregressive Transformer decoder.

In the decoder, we aim to decode the representation of joint coordinates corresponding to joint queries depending on the memory $\mathbf{M}$. The joint queries are learned and transformed through a series of self-attention, cross-attention, and feed-forward networks. Then, the transformed features are fed into a standard MLP head with three fully connected layers to predict 3D joint coordinates. We map the output of each decoder layer to a 3D pose by exploiting the MLP head with shared weights. Similar to HandGCNFormer, joint queries correlate with the hand joints one by one. Therefore, the Hungarian matching algorithm is not necessary for our baseline, which differs from DETR. In our experiments, we adopt four layers for both encoder and decoder as well as apply the same loss function as our HandGCNFormer.

B. Additional Ablation Studies

In this section, we discuss the additional ablation studies on Hands2017. We evaluate HandGCNFormer with the mean of 3D distance error in millimeters. Specifically, "SEEN" and "UNSEEN" indicate the cases whether the test subjects have appeared in the training set. "AVG" denotes the mean of 3D distance error over all subjects. Unless specified, ResNet-50 is used as the backbone, and the size of input cropped images is 256×256.

Auxiliary loss: As mentioned in Section 3.4, we apply an auxiliary loss to supervise the initial pose predicted by a MLP module after ResNet. We quantify the contribution of the auxiliary loss and report results in Table 1. The first model is supervised only by regression loss; the second model is supervised by both regression loss and auxiliary loss. The results show that the second model obtains better accuracy, revealing that the auxiliary loss encourages the ResNet to select more appropriate features and improves the performance.

$\mathcal{L}_{aux}$	$\mathcal{L}_{reg}$	AVG	SEEN	UNSEEN
	✓	6.94	4.72	8.79
✓	✓	**6.80**	**4.64**	**8.59**

Table 1: Ablation study for the effectiveness of auxiliary loss. Best in **bold**.

Backbone	AVG	SEEN	UNSEEN	Params(Flops)
ResNet-18	7.31	5.03	9.22	**19.93M(2.7G)**
ResNet-50	6.80	**4.64**	8.59	33.04M(5.7G)
ResNet-101	**6.78**	4.66	**8.55**	52.04M(10.6G)

Table 2: Ablation study for the effectiveness of different backbone. Best in **bold**.

Backbone: The results of our model with different backbones are reported in Table 2. The larger backbone improves overall performance. In particular, our method with the ResNet-18 surpasses AWR [3] yet only contains 58.6% parameters of it. AWR with ResNet-50 has 34M parameters.

Input size: Table 3 reports the performance of HandGCNFormer with different input image sizes. The results demonstrate that the larger input image achieves better accuracy because it encourages ResNet and Transformer encoder modules to extract more fine-grained context infor-

Input Size	AVG	SEEN	UNSEEN	Params(Flops)
192×192	7.12	4.85	8.73	**31.21M(3.21G)**
256×256	6.80	4.64	8.59	33.04M(5.70G)
320×320	**6.72**	**4.59**	**8.50**	35.40M(8.92G)

Table 3: Ablation study for the effectiveness of different input image sizes. Best in **bold**.

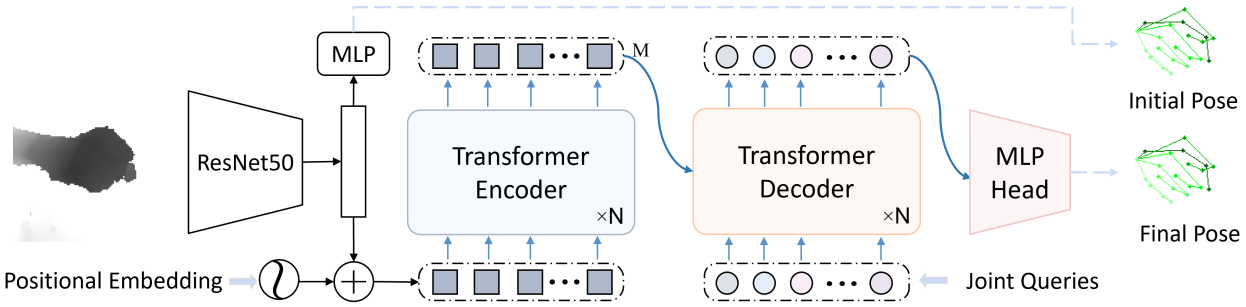

Figure 1: The overview of our baseline.

mation, providing more visual evidence for inferring joint locations.

C. Additional Visualization Results

Figures 2 and 3 show the qualitative results of HandGCNFormer in Hands2017 dataset under severe self-occlusion and high self-similarity cases, respectively. Benefiting from the synergy of Transformer and GCN, our method achieves excellent performance. For invisible joints that lack sufficient local depth information and similar joints that exhibit analogous local appearance, our method can infer plausible locations based on both global context information and hand kinematic topology.

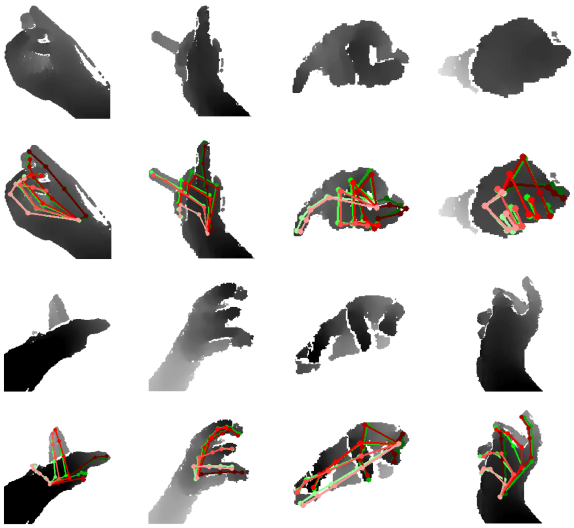

Figure 2: Qualitative results of images with self-occlusion on Hands2017 dataset. Red pose represents the ground truth. Green pose is predicted result. The first and third rows are the depth images, and the rest are the corresponding results of HandGCNFormer.

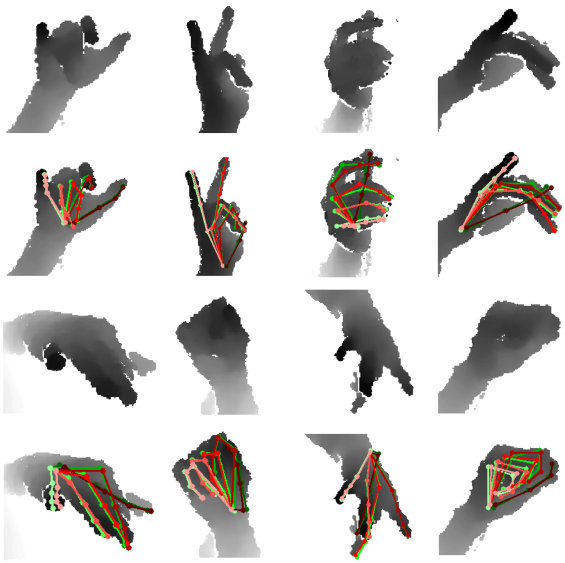

Figure 3: Qualitative results of images with self-similarity on Hands2017 dataset. Red pose represents the ground truth. Green pose is predicted result. The first and third rows are the depth images, and the rest are the corresponding results of HandGCNFormer.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[3] Weiting Huang, Pengfei Ren, Jingyu Wang, Qi Qi, and Haifeng Sun. Awr: Adaptive weighting regression for 3d hand pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11061–11068, 2020. 1