A Additional Network Details

A.1 Spiral Operator

In our mesh decoder, we use the spiral operator (Gong et al., 2019) to process vertex features in the spatial domain. The spiral operator defines a spiral selection of neighbors for a center vertex v by imposing the order on elements of the $k-$disk. A $k-$ring and a $k-$disk around a center vertex v can be defined as follows:

$$\begin{aligned} 0-ring(v) &= v, \\ k-disk(v) &= \bigcup_{i=0\ldots k} i-ring(v), \\ (k+1)-ring(v) &= \mathcal{N}(k-ring(v))\backslash k-disk(v). \end{aligned} \tag{1}$$

where $\mathcal{N}(\mathcal{V})$ denotes the vertex neighborhood.

The spiral length is denoted as l, then we can get an ordered set $O(v,l)$ consisting of l vertices by concatenating of $k-$rings:

$$O(v,l) \subset (0-ring(v), 1-ring(v), ..., k-ring(v)). \tag{2}$$

After getting the sequence, we can define the convolution in the manner of the euclidean convolution. The spiral convolution operator for a node i can be defined as:

$$G^i_{s+1} = \gamma_{s+1}(\|_{j\in S(i,l)}\ G^j_s) \tag{3}$$

where γ is MLPs and $\|$ denotes the concatenation operation.

A.2 Mesh Sampling

We apply a multi-scale hand mesh representation to capture both global and local information. At each stage of the mesh decoder, the number of mesh vertices are changed by the factor of 2. We follow COMA (Ranjan et al., 2018) to perform the in-network sampling operations(down-sampling and up-sampling) using pre-computed transform matrices. The down-sampling matrix D is obtained by iteratively contracting mesh vertex pairs while maintaining surface error approximation using quadric metrics. The up-sampling matrix U is obtained by including barycentric coordinates of the vertices which are discarded during the downsampling.

Table 1: Analysis on different backbones, evaluated on FreiHAND. All backbones are pre-trained on ImageNet.

Backbone	PA-MPJPE↓	PA-MPVPE↓
HRNet-W64	6.4	6.5
ResNet-50	6.7	6.9

A.3 Network Architecture details

For the feature encoder, we adopt different backbones including HRNet-W64 (Wang et al., 2020) and ResNet-50 (He et al., 2016). The backbone is followed by a series of deconvolution layers to obtain multi-scale feature maps F_s and H_s. The feature maps of the same stage have the same channels and sizes *i.e.* 2048×7×7, 1024×14×14, 512×28×28, 256×56×56. We employ four 1×1 convolution layers to get $Q \in \{512\times7\times7, 256\times14\times14, 128\times28\times28, 64\times56\times56\}$. The feature maps are then passed to four mesh decoding blocks. The output channels of each block are the same as Q size in each stage s. A block consists of a spiral convolution layer and a self-attention module. For the spiral convolution, we set the spiral length l as 27. For the self-attention module, we set the number of heads to 4.

We study the behavior of different encoder backbones. Both HRNet (Wang et al., 2020) and ResNet (He et al., 2016) model are pre-trained on the ImageNet. In table 1, we find that HRNet has greater performance than using ResNet-50. As HRNet uses high-resolution feature pyramids, we observe further improvement.

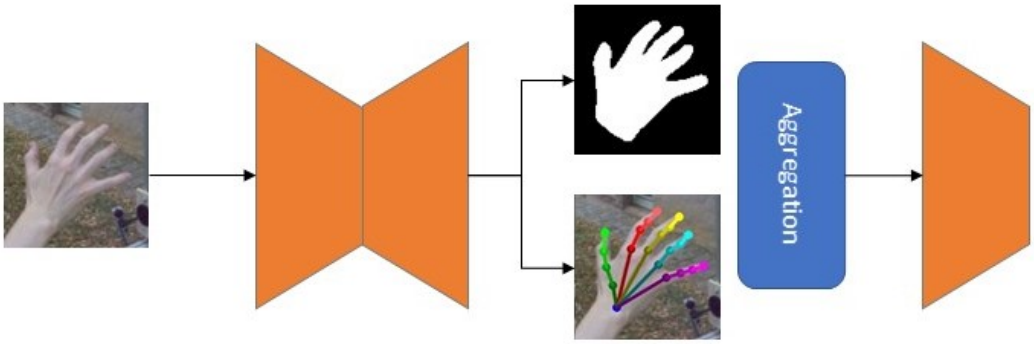

Figure 1: The re-designed 2D image feature extractor.

A.4 THE 2D FEATURE EXTRACTOR FOR GCN METHOD

Inspired by (Chen et al., 2021), we propose to add semantic aggregation to our 2D feature extractor. Specifically, we first utilize an image encoder-decoder structure that predicts 2D joint landmarks and masks. Then, we combine all the predicted joint heatmaps to aggregate 2D joint locations, which can take advantage of our 2D auxiliary tasks. Then, we use another encoder backbone to encode the combined heatmaps and exploit the feature maps as previous settings. Figure 1 shows the pipeline.

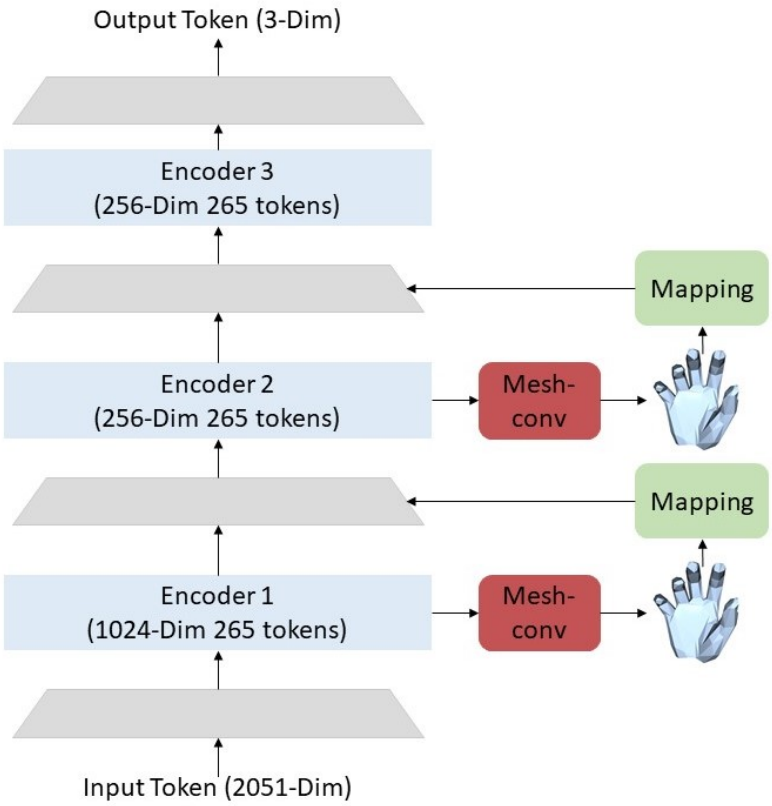

Figure 2: Architecture of transformer-based method.

B ARCHITECTURE OF TRANSFORMER-BASED METHOD

We inject our feature mapping and the hierarchical decoding into the Multi-Layer transformer encoder (Lin et al., 2021). Given the joint and mesh vertex queries, the encoder maps the input to 3D hand joints and mesh vertices.

As shown in Figure 2, the encoder consists of three blocks. All three blocks have the same number of tokens. The hidden dimensions are 1024, 256, and 64. For the first two blocks, we use two additional mesh-conv layers to predict intermediate mesh results. The pixel-aligned features with the same dimension as blocks are obtained using the mapping module and then concatenate to mesh features. At the output of each encoder block, we use the MLP layers for dimension reduction.

C ADDITIONAL RESULTS

Figure 3 illustrates the comprehensive qualitative results of our method. The challenges of the input include complicated hand poses and hand-object interaction situations. Overcoming these difficulties, our method can predict accurate hand mesh as well as silhouette and 2D pose.

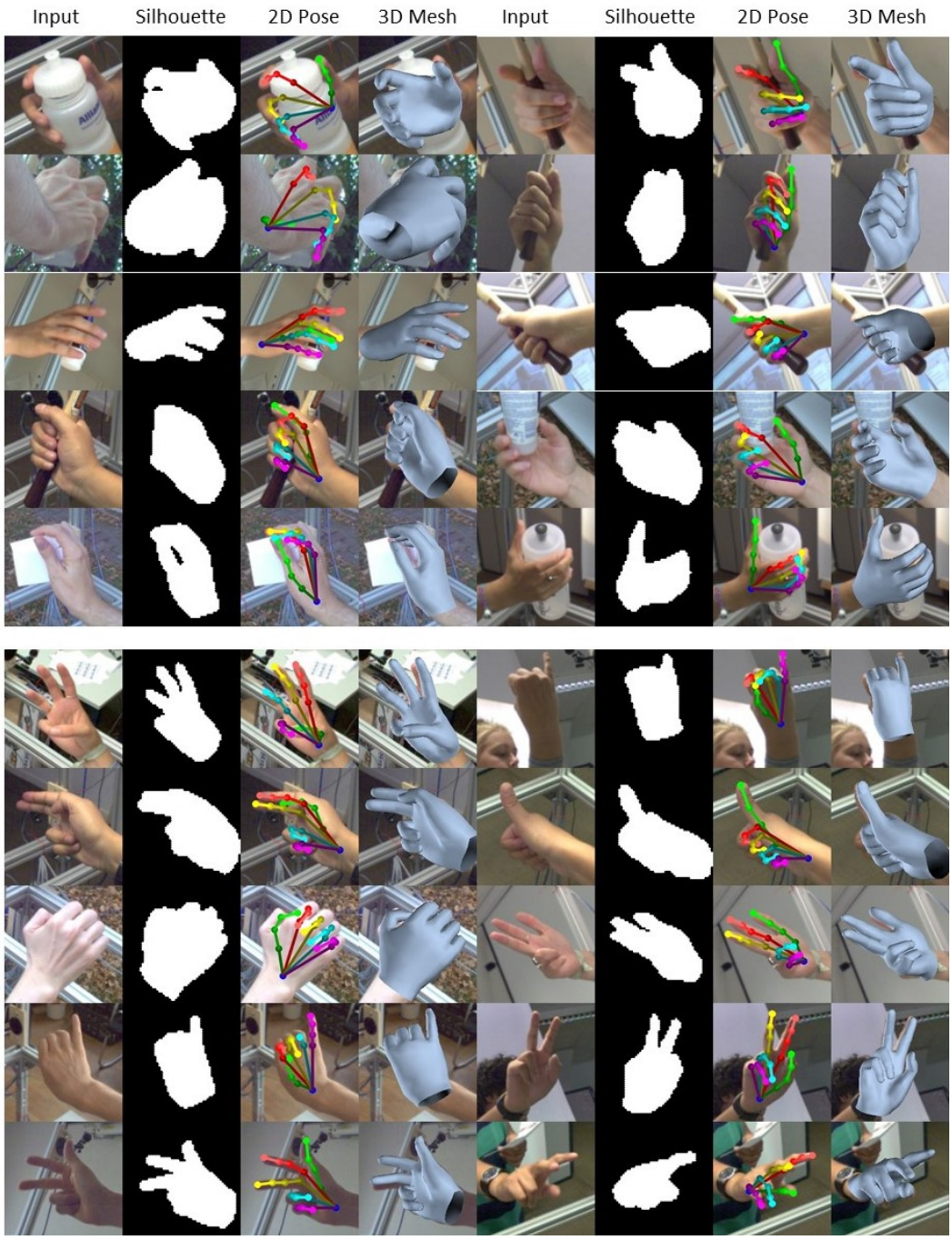

Figure 3: Qualitative results on FreiHAND dataset.

REFERENCES

Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13274–13283, 2021.

Shunwang Gong, Lei Chen, Michael Bronstein, and Stefanos Zafeiriou. Spiralnet++: A fast and highly efficient mesh convolution operator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

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*, pp. 770–778, 2016.

Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1954–1963, 2021.

Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 704–720, 2018.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.