Supplementary Materials for High Fidelity 3D Hand Shape Reconstruction via Scalable Graph Frequency Decomposition

Tianyu Luan[1] Yuanhao Zhai[1] Jingjing Meng[1] Zhong Li[2]
Zhang Chen[2] Yi Xu[2] Junsong Yuan[1]

[1]State University of New York at Buffalo [2]OPPO US Research Center, InnoPeak Technology, Inc

1. Detailed Network Architecture

We proposed a detailed network architecture of our approach in Fig. 1. The green boxes are the features, in which we note the feature dimensions. The blue boxes represent blocks of EfficientNet [3]. The red boxes represent GCN blocks. The GCN residual blocks in the network are designed following the manner of [1]. Details of the residual blocks are shown on the right of the figure. The gray boxes are the feature skip-connection part. To get multi-level image features from feature maps, we project the vertices into the feature maps, and use a bilinear interpolation technique to sample features. We will illustrate the process more in Sec. 2. The purple boxes are the sub-network used to generate MANO mesh. The orange boxes indicate the annotation we used. The green arrows are feature streams and the red lines are skip connections.

We fetch skip-connected features from the output of EfficientNet Block 1, Block 3, and Block 7. The features are used as parts of the input of the GCN. The GCN has 3 levels. At each level, the input features go through a 10-layer GCN Residual Block, then output a feature vector and a 3D location at each vertex. The 3D locations are used as intermediate output and for supervision. The features are used as a part of the input for the next level. At the third level, we only output the 3D location of each vertex as the final mesh.

2. Skip-connected Feature Sampling

In Fig. 1, the features fetched from EfficientNet are feature maps. We want to transfer them into feature vectors and put them on the vertices without losing spacial information. Thus, we design a feature sampling strategy to put the local image feature on each graph vertex. As shown in Fig. 2, we use orthodox projection to find the feature vector for each vertex on the feature map. For every vertex P, we calculate the projection point P' on the feature map. Then, we extract the feature vector $x \in \mathbf{R}^c$ using bilinear interpolation at point P', where c is the feature map channel number. The total output feature dimension is $N \times c$, where N is the number of graph vertices.

3. Mesh Post-processing

We do a post-process on the third-level mesh. Due to the flaws of groundtruth mesh (shown in Fig. 4), some of our output mesh also have similar structure flaws. To tackle this problem, we designed a smooth mask to reduce the flaws. Fig. 3 shows the output of the network, our smooth mask, and our final mesh result. As we can see, the flaws are highly reduced. Note that, this flaw is caused by the noisy groundtruth, so it can also be reduced by a better remeshing of the training data in the future.

4. Remeshing Procedure

We try to use the multiview stereo (MVS) generated mesh provided in [2]. However, the MVS mesh has about 500k vertices on each mesh. The large vertex number mesh with high redundancy makes our training process much slower. Moreover, without a fixed topology, the choices of shape supervision are limited. For example, we would not be able to use the per vertex loss and frequency decomposition loss for training.

Thus, we designed a remeshing technic to transfer the mesh generated in the multiview stereo (MVS) method into a unified topology. The algorithm is shown in Fig. 4a. First, we align the MVS mesh with a parametric template mesh. Here, we use template meshes designed in the main paper Section 3.2. Second, we use an optimization approach to calculate a set of pose and shape parameters, so that the template mesh becomes a coarse approximation of the MVS mesh. Finally, we use the closet point on the MVS mesh as a substitute for each vertex on the parametric mesh. This procedure would preserve the detailed shape and the topology of the parametric template at the same time. In our experiments, we generate 3 resolution levels of groundtruth mesh for

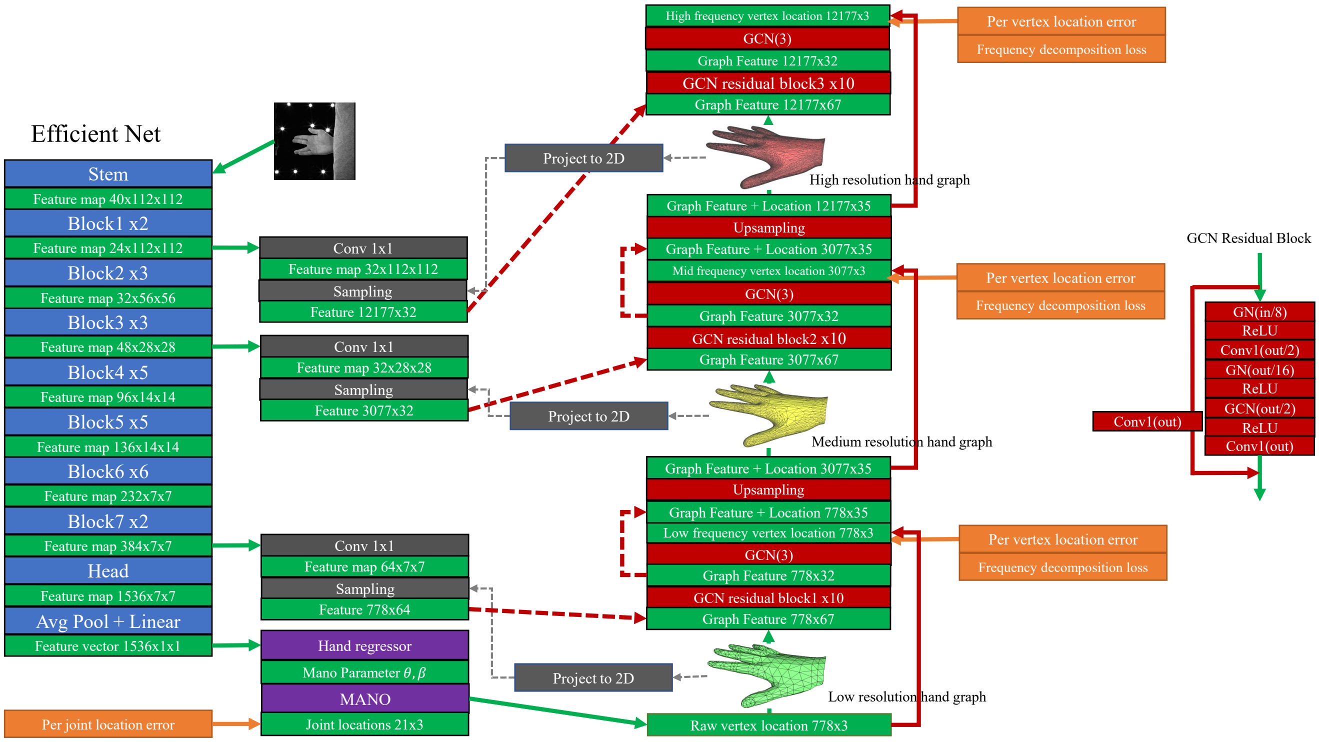

Figure 1. The detailed network architecture.

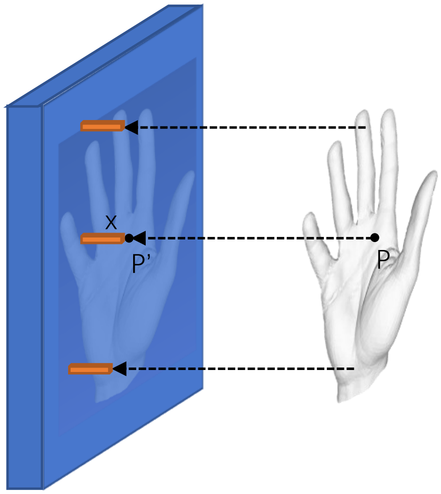

Figure 2. Skip-connected Feature Sampling.

supervision, and use the third level for testing.

However, despite the good attributes of the groundtruth meshes, some of them still have flaws. Fig. 4b shows an example of the mesh flaws inside the mesh (red rectangle). It happens because some of the vertices on the parametric mesh find the wrong corresponding vertices on the MVS mesh. These groundtruth mesh flaws will eventually cause defects on generated mesh (shown in Fig. 3). We have largely reduced the flaws of our mesh using the mesh post-processing method mentioned in Sec. 3.

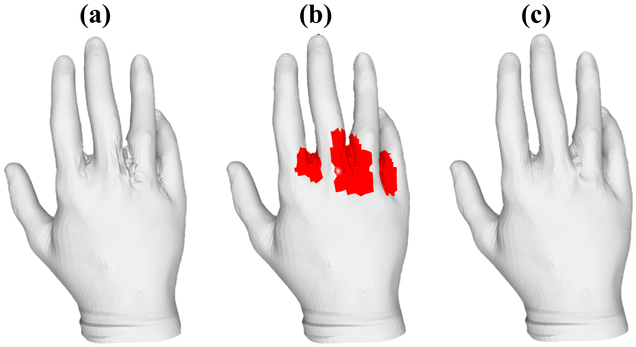

Figure 3. Mesh pose-processing. a) Original mesh. b) Smoothing mask (Red). c) Final result.

5. More Visualization Results

We show more visualization results of our proposed method in Fig. 5.

6. Failure Cases and Future Works

We show in Fig. 6 a few failure cases where our method generates hand meshes with flaws. Most of these flaws are caused by groundtruth flaws in remeshing (shown in Fig. 4b).

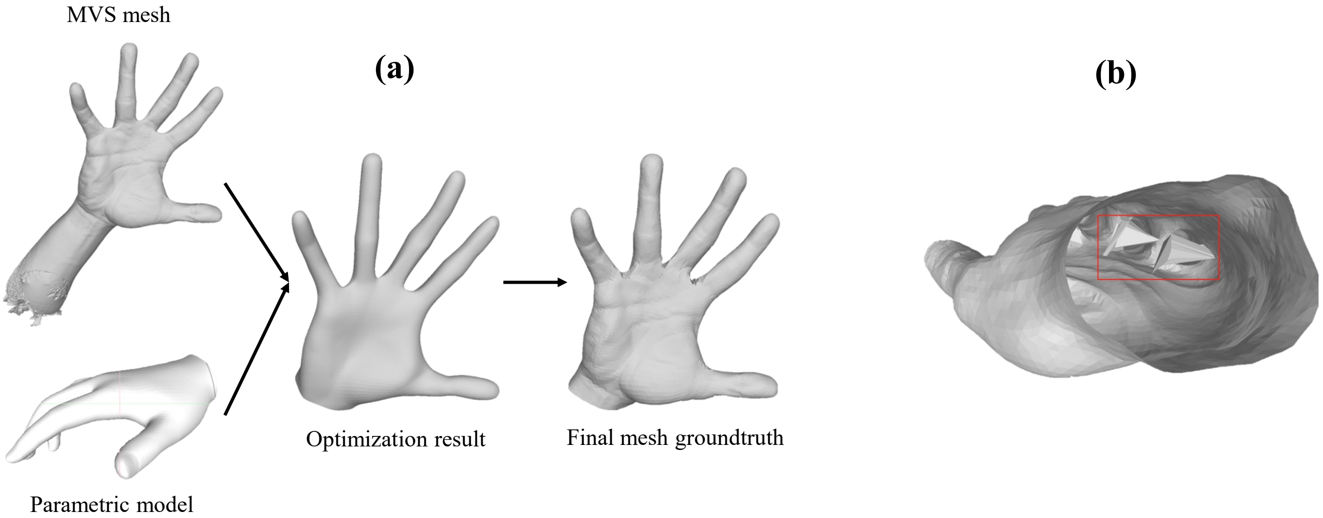

Figure 4. a) Remeshing procedure. b) Example of groundtruth flaws

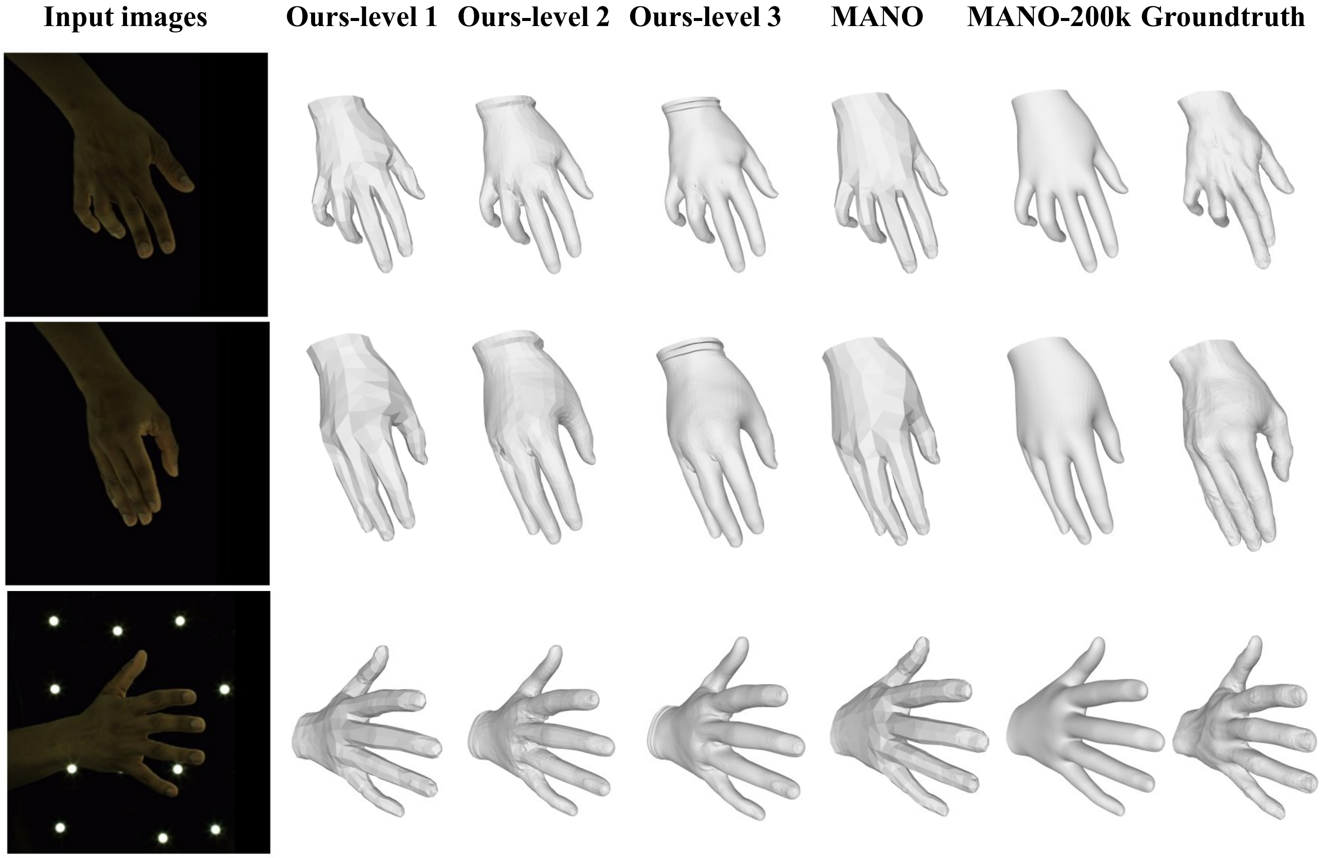

Figure 5. More visualization results.

In future works, we would improve the remeshing procedure to reduce the artifacts. Besides, we would also improve our method to tackle the in-the-wild hand reconstruction problem.

References

[1] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, pages 4501–4510, 2019. 1

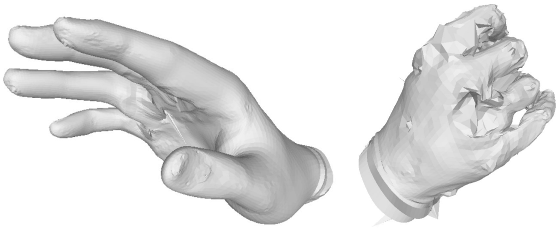

Figure 6. Failure cases.

[2] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *ECCV*, 2020. 1

[3] Mingxing Tan and Quoc V Le. Efficientnet: Improving accuracy and efficiency through automl and model scaling. *arXiv preprint arXiv:1905.11946*, 2019. 1