I2UV-HandNet: Image-to-UV Prediction Network for Accurate and High-fidelity 3D Hand Mesh Modeling *(Supplementary Material)*

Ping Chen[1] Yujin Chen[2,3] Dong Yang[1] Fangyin Wu[1] Qin Li[1] Qingpei Xia[1] Yong Tan[1]
[1]IQIYI Inc. [2]Wuhan University [3]Technical University of Munich
{redcping, terencecyj}@gmail.com {yangdong01, wufangying, liqin01, xiaqingpei, tanyong}@qiyi.com

In this appendix, we detail our network architecture in Section A; in Section B, we additionally provide more results.

A. Network Architecture Details

We detail the network architecture in detail in Tables 8-10. Table 8 describes the size of each feature map from the encoder. Here, we use five feature maps of ResNet-50 with a convolution operation using 3×3 convolution kernels and the output channels are 64, 128, 256, 512, and 1024. The feature maps are then fed to the decoder of the proposed AffineNet which predicts UV position maps in multiple resolutions, as detailed in Table 9. We note that "affine-operation" indicates an affine transformation from E^i to A^i according to the current UV map prediction I^i_{UV}. Along the expansive path, the UV position map gets more and more accurate (from the coarsest I^4_{UV} to the most accurate I^0_{UV}), thus the A^i gets better accurate spatial alignment, then resulting in more accurate predicted UV position map of the next level I^{i-1}_{UV}. In conclusion, the prediction of the UV position map and the degradation of coordinate ambiguities are coarse-to-fine and mutually improved. The SRNet is described in Table 10.

B. Additional Results

We show additional evaluate on the impact of E_{grad} in Figure 5 (also see Section 4.6 of the main text). As shown in the figure, E_{grad} makes the mesh have better surface continuity, that is, smoother. The two samples that do not use E_{grad} have unreasonable folds or depressions on their surfaces. We show additional visualization of our predictions via AffineNet and SRNet for FreiHAND test set [3] (see Figure 6), HO3D test set [1] (see Figure 7), and HIC dataset [2] (see Figure 8). Here, the AffineNet is trained on a combination of data (the training data of FreiHAND, Obman, and YT-3D) and the SRNet is trained on the SHS dataset. The visualization shows our method outputs accurate and high-fidelity hand reconstruction results, and the results on HO3D and HIC show that our network has good generalization.

Even if the method is generally robust, we find there may exists unreasonable mesh collapse when the hand is cropped outside the image scope.

The whole network runs 46fps on an NVIDIA Tesla V100. For the VR application, we implemented the model on the Qualcomm Snapdragon XR2 platform, replacing MobileNet-v3 as the encoder, resulting in a quantized model of 19MB and running above 50fps.

References

[1] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[2] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118(2):172–193, 2016.

[3] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *International Conference on Computer Vision*, 2019.

Encoder Feature Map	Size
Input	(3,256,256)
E^1	(64,128,128)
E^2	(128,64,64)
E^3	(256,32,32)
E^4	(512,16,16)
E^5	(1024,8,8)

Table 8: The size of each feature map in the encoder.

Input	Operation	Output	Output Size
E^5	Upsample/Conv/BN/ReLU	D^4	(512,16,16)
D^4	Conv/Sigmoid	I_{UV}^4	(3,16,16)
E^4,I_{UV}^4	Affine-operation/Upsample	A^3	(512,32,32)
D^4	Upsample/Conv/BN/ReLU	D^3	(256,32,32)
I_{UV}^4	Upsample	$\hat{I}_{UV}^3$	(3,32,32)
A^3,D^3,$\hat{I}_{UV}^3$	Concat/Conv/BN/ReLU	D'^3	(256,32,32)
D'^3	Conv/Sigmoid	I_{UV}^3	(3,32,32)
E^3,I_{UV}^3	Affine-operation/Upsample	A^2	(256,64,64)
D'^3	Upsample/Conv/BN/ReLU	D^2	(128,64,64)
I_{UV}^3	Upsample	$\hat{I}_{UV}^2$	(3,64,64)
A^2,D^2,$\hat{I}_{UV}^2$	Concat/Conv/BN/ReLU	D'^2	(128,64,64)
D'^2	Conv/Sigmoid	I_{UV}^2	(3,64,64)
E^2,I_{UV}^2	Affine-operation/Upsample	A^1	(128,128,128)
D'^2	Upsample/Conv/BN/ReLU	D^1	(64,128,128)
I_{UV}^2	Upsample	$\hat{I}_{UV}^1$	(3,128,128)
A^1,D^1,$\hat{I}_{UV}^1$	Concat/Conv/BN/ReLU	D'^1	(64,128,128)
D'^1	Conv/Sigmoid	I_{UV}^1	(3,128,128)
E^1,I_{UV}^1	Affine-operation/Upsample	A^0	(64,256,256)
D'^1	Upsample/Conv/BN/ReLU	D^0	(32,256,256)
I_{UV}^1	Upsample	$\hat{I}_{UV}^0$	(3,256,256)
A^0,D^0,$\hat{I}_{UV}^0$	Concat/Conv/BN/ReLU	D'^0	(32,256,256)
D'^0	Conv/Sigmoid	I_{UV}^0	(3,256,256)

Table 9: Layer specification for the decoder part. "Upsample" indicates to use bilinear interpolation to enlarge the spatial size by 2 times; "Conv" indicates convolution operation using 3×3 convolution kernel with zero padding; "BN" indicates batch normalization.

Layer	Operation	Input Size	Output Size	Kernel Size	Padding
Level 1	Conv/ReLU	(3,256,256)	(64,256,256)	(9,9)	(4,4)
Level 2	Conv/ReLU	(64,256,256)	(32,256,256)	(5,5)	(2,2)
Output	Conv	(32,256,256)	(3,256,256)	(5,5)	(2,2)

Table 10: Layer specification for the SRNet.

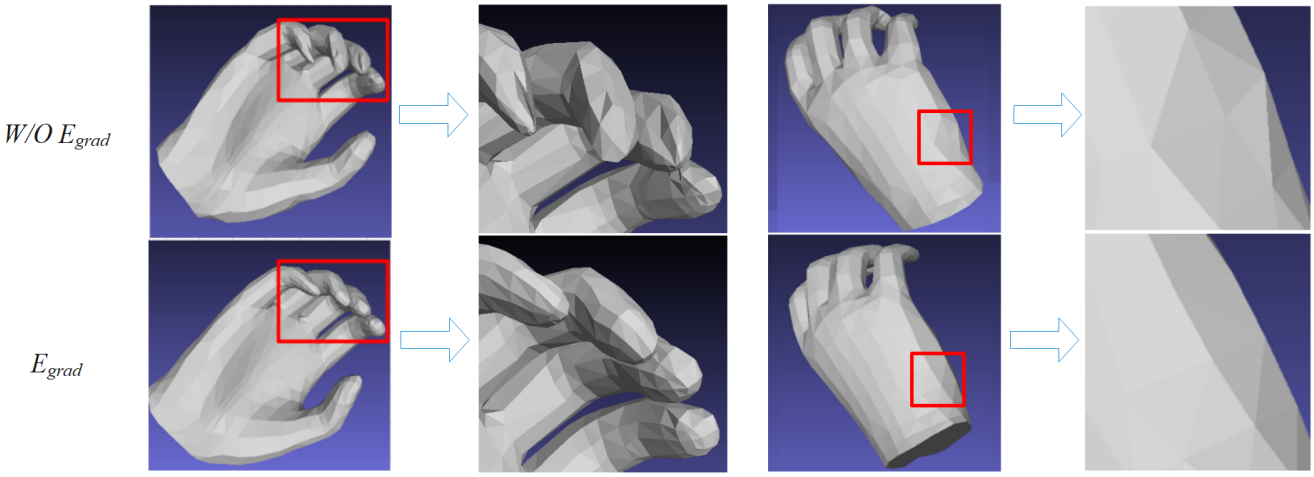

Figure 5: Qualitative comparison of 3D hand reconstruction results from AffineNet in terms of E_{grad} used or not.

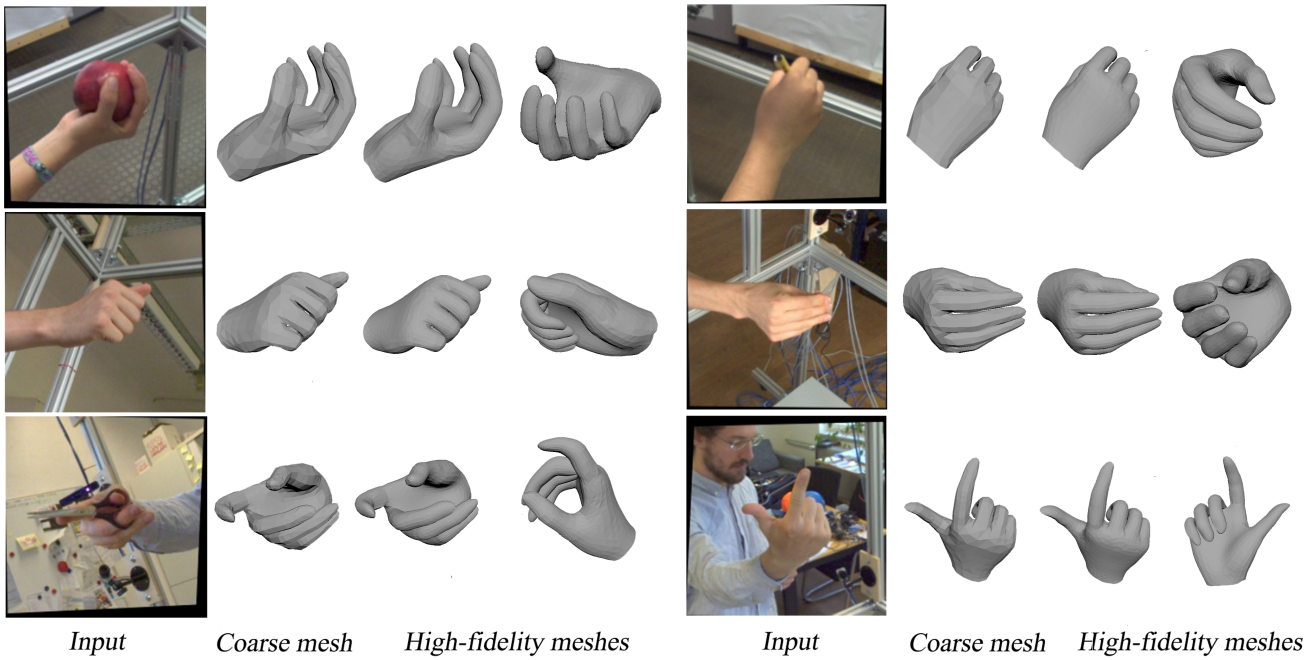

Figure 6: Qualitative visualization of our method on the FreiHAND testing set.

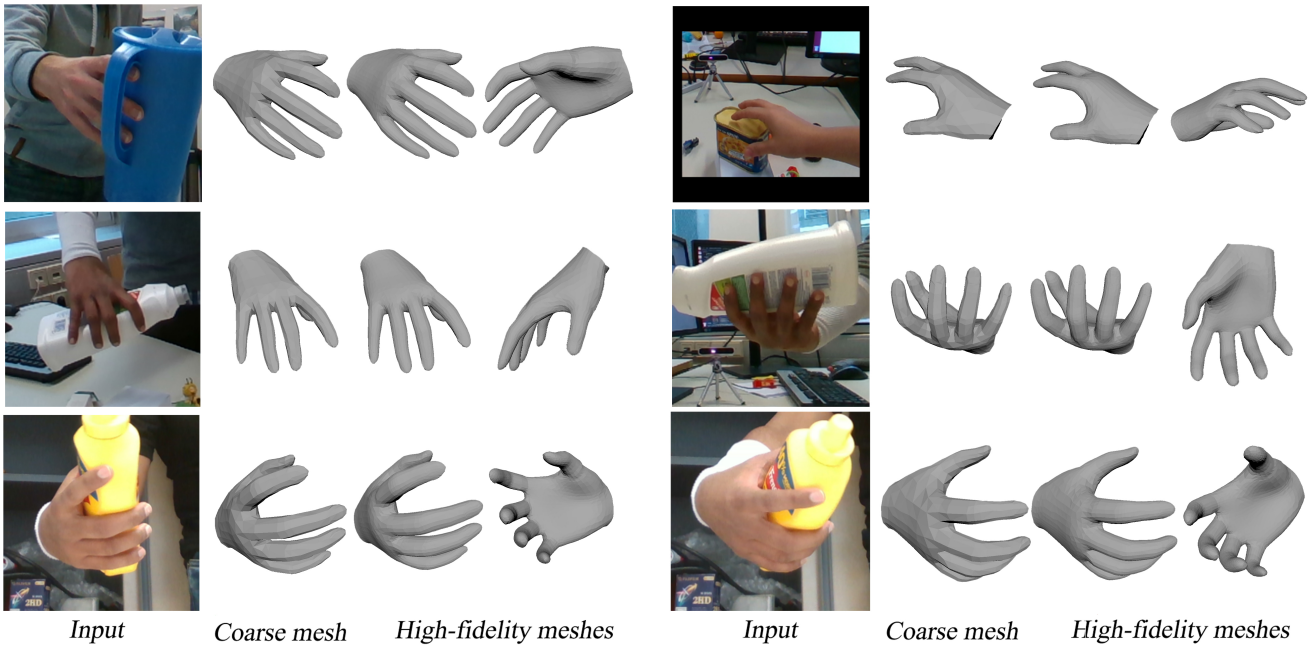

Figure 7: Qualitative visualization of our method on the HO3D testing set.

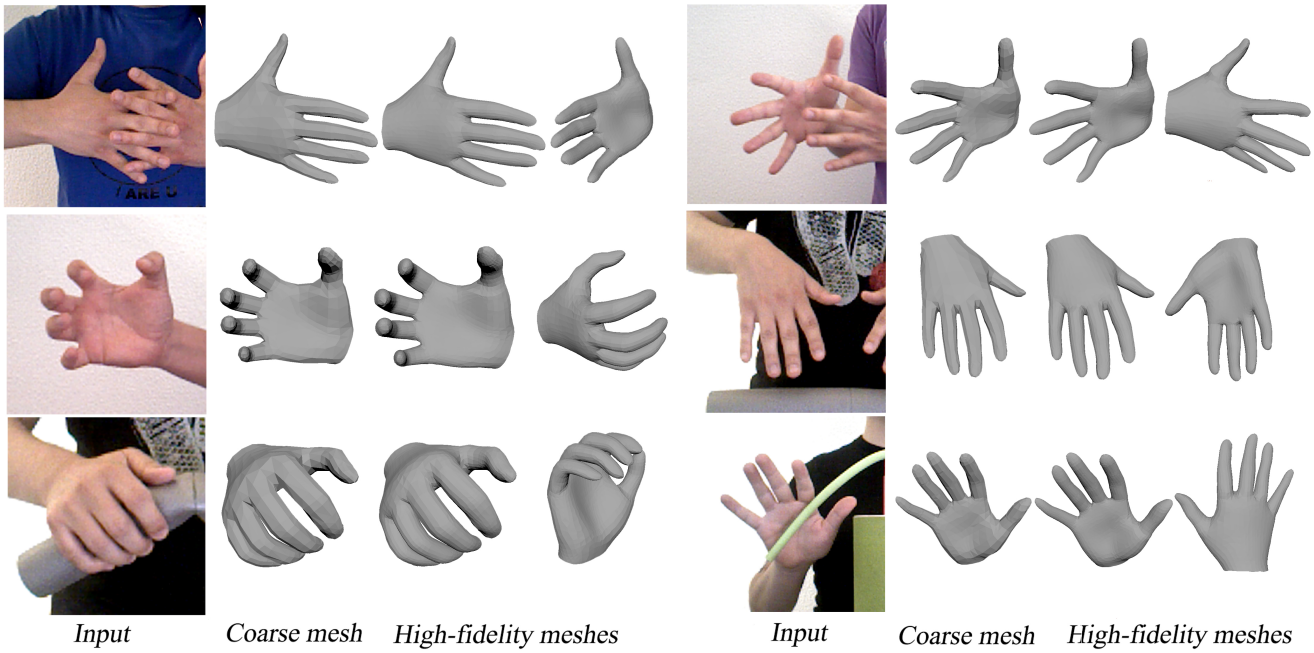

Figure 8: Qualitative visualization of our method on the HIC dataset.