Supplementary Material
Mesh Graphormer

Kevin Lin Lijuan Wang Zicheng Liu
Microsoft
{keli, lijuanw, zliu}@microsoft.com

Figure 1: Qualitative results when there are heavy occlusions. For each example, we show results from METRO [3] and Graphormer. We can see that both METRO and Graphormer are quite robust against occlusions, but Graphormer generates more favorable head pose and body pose. Blue: METRO. Silver: Graphormer.

A. Qualitative Comparison

Figure 1 shows qualitative results of Graphormer compared with METRO [3] in the scenario of heavy occlusions. We can see that both methods are quite robust against occlusions, but Graphormer generates better head and body poses. At the top right of Figure 1, almost half of the subject is occluded. Graphormer reconstructs a human mesh with more accurate head/body pose compared to METRO. At the bottom right of Figure 1, the subject is occluded by the car door. We see Graphormer reconstructs a more reasonable body shape. At the bottom left, the subject is standing behind the fence. Our method reconstructs a human mesh with the two legs better aligned with the image. The results demonstrate the effectiveness of the proposed method.

B. Additional Qualitative Results

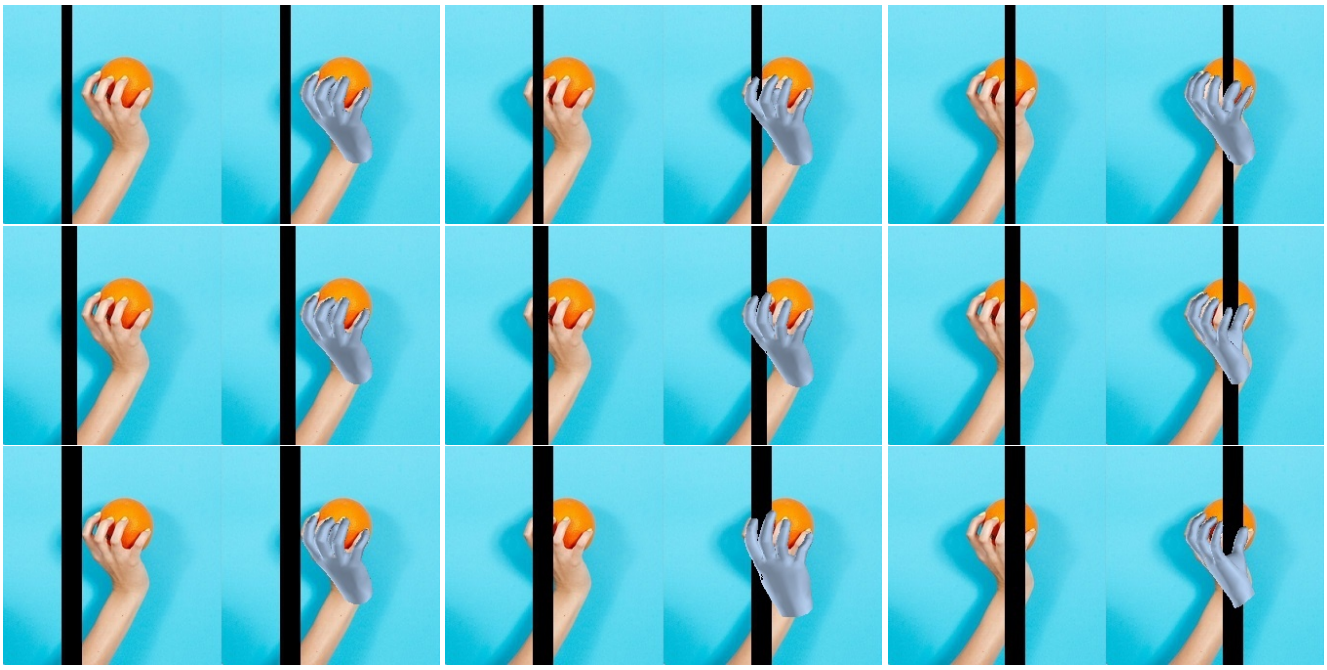
Further, to demonstrate the robustness and generalization capability of our model to challenging scenarios, we test our model on the hand images that are collected from the Internet. The images have severe occlusions with different objects.

To make the task even more difficult, we create artificial occlusions including black vertical stripes to cover one or two fingers, or part of the palm of the hand in the test images. Please note that the artificial occlusions are only used in the inference stage. We do not use any artificial occlusions in training.

In Figure 2, Figure 3, Figure 4, and Figure 5, we show the input images and our reconstructed hand meshes. For each figure, the top row shows the occlusion scenario with narrow black stripes. From the second row to the bottom, we gradually increase the width of the stripes to occlude more fingers or more parts of the hand.

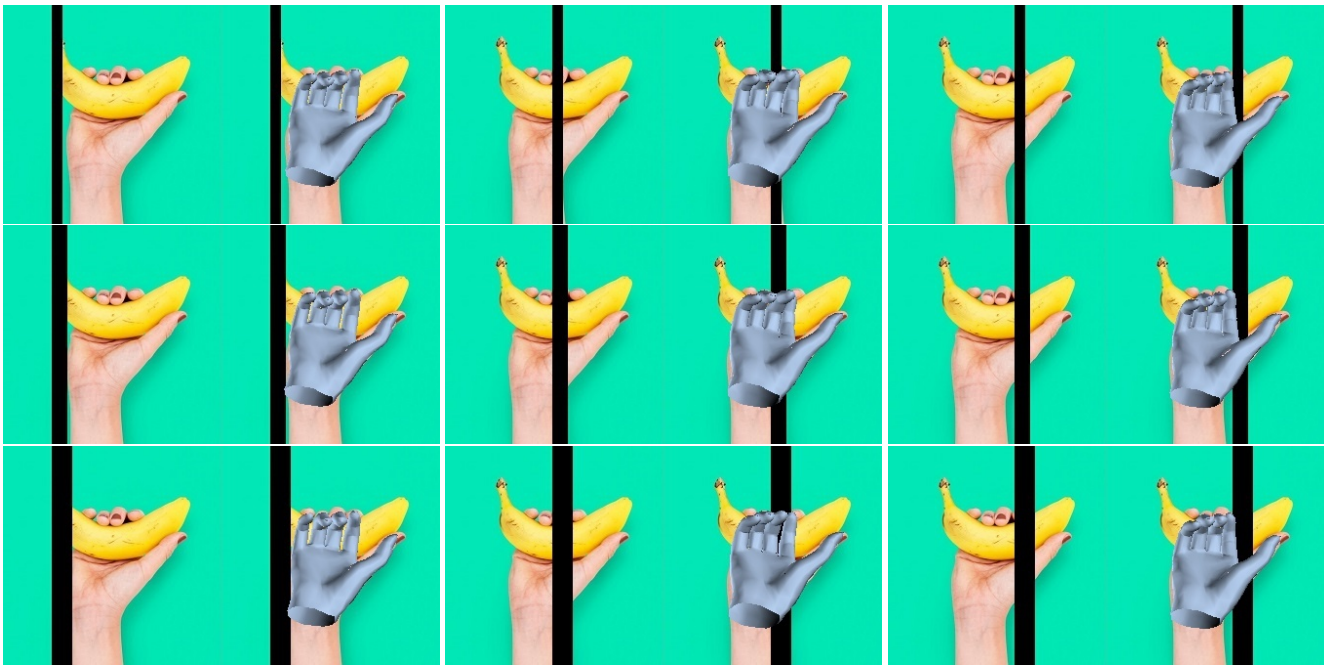
Figure 2 shows a hand with an orange. Our model is able to reconstruct a reasonable hand mesh, even if the hand is severely obscured by the vertical stripes. The results show that Graphormer is to some extent robust to the artificial occlusion patterns.

In Figure 3, there is a hand grasping a banana. Although the banana is a novel object unseen in training and



Input Output Input Output Input Output

Figure 2: Qualitative results of our method. There is a hand with an orange. We demonstrate the robustness of our model by adding artificial occlusions including black vertical stripes to the images. We can see that Graphormer reconstructs plausible hand mesh under the occlusion scenarios. Please see `Fig1.gif` for more detailed video results.



Input Output Input Output Input Output

Figure 3: Qualitative results of our method. There is a hand holding a banana. We do not have any banana training images. However, Graphormer generalizes to the novel object, and creates the hand mesh with the correct pose. Please see `Fig2.gif` for more detailed video results.

Figure 4: Qualitative results of our method. There is a hand holding an ice cream cone. The ice cream cone is a novel object unseen in training, and the hand pose is also object-specific. We can see that Graphormer works reasonably well for the test images. Please see `Fig3.gif` for more detailed video results.

Method	Para (M)	Δ Para (M)	PAMPJPE ↓
Graphormer - GRB	98.39	–	35.9
Graphormer - GRB + MLP1	98.43	0.04	35.9
Graphormer - GRB + MLP2	98.92	0.53	36.0
Graphormer	98.43	0.04	34.5

Table 1: Comparison between the use of MLP and GRB.

a large portion of the fingers are occluded by the banana, Graphormer successfully reconstructs the hand mesh under various occlusion scenarios. This demonstrates the generalization ability of our proposed Graphormer.

Figure 4 shows a hand with an ice cream cone. Please note that the ice cream cone is unseen during training, and the interaction between the hand and the ice cream cone is complex. However, our model generalizes well to the test images. Even though the occlusions by the ice cream cone are severe and sometimes most of the fingers are occluded by the black vertical stripes, Graphormer still reconstructs a reasonable hand shape with object-specific grasp.

Figure 5 shows a hand with half an orange. Most of the fingers are invisible in this image. However, our model creates a hand mesh with the correct hand pose.

We further present the video results of Figure 2, Figure 3, Figure 4, and Figure 5. Please find the video results in the attached GIF files.

C. Comparison between MLP and GCN

In this section, we replace graph residual block with MLPs, and study the performance of the use of MLPs with a similar or larger model size.

In Table 1, the first row corresponds to the baseline transformer that uses image grid features. In the second and the third rows, we gradually increase the hidden size of the MLP module in the transformer, but we do not achieve any gain in performance. The bottom row of Table 1 shows the results of our Graphormer. As can be seen, adding graph convolutions has a slight increase of 0.04M parameters, but it improves performance significantly from 35.9 to 34.5 PA-MPJPE.

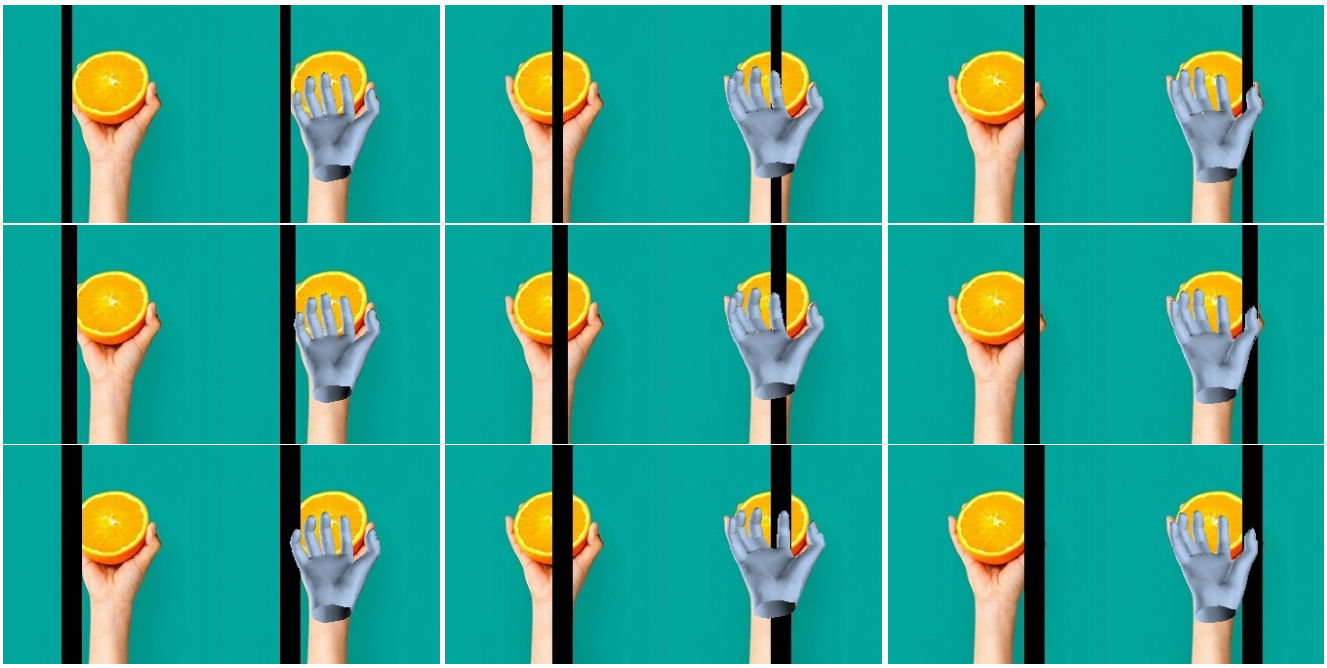


Figure 5: Qualitative results of our method. It is a hand holding half an orange. Graphormer is able to reconstruct a reasonable hand mesh even though most of the fingers are invisible. Please see `Fig4.gif` for more detailed video results.

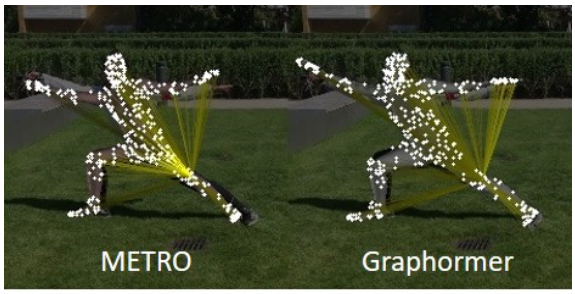


Figure 6: Attention map without color normalization. Graphormer pays more attentions to the lower left leg compared to METRO.

D. Design Options of Graphormer Encoder

In Figure 7, we graphically illustrate three design options of the Graphormer encoder we have studied in the paper. Please refer to Table 6 in our main paper for the performance comparisons between the design options. We observe that placing graph convolutions after MHSA works better than other design options for the reconstruction of human mesh.

E. Discussion of Attention Map

Please note that the attention colors in the paper's diagrams are normalized based on the maximum attention value. Because the maximum attention value for Graphormer is smaller, so the overall colors are lighter. We attach the two diagrams without color normalization in Figure 6. We see that both methods pay similar attention on the left arm and right foot, while Graphormer also attends to the left lower leg.

F. Discussion of Camera Parameters

We learn camera parameters for a weak perspective camera model. Following [2], we predict a scaling factor s and a 2D translation vector t. Please note that the model prediction is already in the camera coordinate system, thus we don't have to compute global camera rotation. The camera parameters are learned via 2D pose re-projection optimization. It doesn't require any GT camera parameters.

G. Training Time

We conducted experiments on a machine with 8 NVIDIA V100 GPUs. We use a batch size of 32. For each epoch, our training takes about 35 minutes. We train the proposed model for 200 epochs. The overall training takes 5 days.

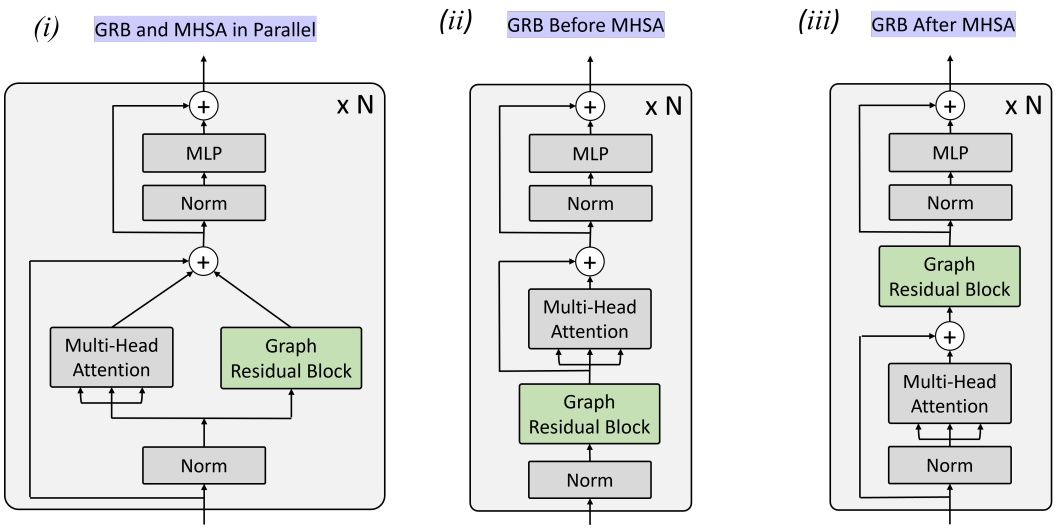


Figure 7: Three design options we have studied for building our proposed Graphormer Encoder. The designs are inspired by language and speech literature [1, 4].

	HRNet	Transformer	Graphormer
# Parameters (M)	128.05	98.39	98.43
GFLOPs	28.89	27.71	27.72

Table 2: Number of parameters and computational complexity in terms of GFLOPs.

H. Computational Costs

Since we inject graph convolutions into the transformer, one may wonder about the computational costs of the proposed Graphormer. To answer the question, we report the number of parameters and the computational complexity in terms of GFLOPs.

Table 2 shows the comparison between the conventional transformer and the proposed Graphormer. We also report the computational cost of the HRNet CNN backbone for reference. As we can see, adding graph convolutions has a slight increase of 0.04M parameters and 0.01 GFLOPs compared to the conventional transformer. The results suggest that little complexity has been added to the transformer architecture. However, Graphormer significantly improves the state-of-the-art performance across multiple benchmarks. This verifies the effectiveness of the proposed method.

Please note that the total parameters of our end-to-end pipeline is the sum of HRNet and Graphormer.

I. Limitation

We observed that our method may not work well if the reconstruction target is out of the view. For example, as shown in Figure 8(a), when the majority of the human body is not in the input image, our method fails to estimate a correct human mesh. This is probably due to the lack of out-of-the-view 3D training data in our training set. In Figure 8(b), only two hands are visible and the rest of the human body is out of the view. Our method does not work well in this case. We plan to address this issue in our future work.

References

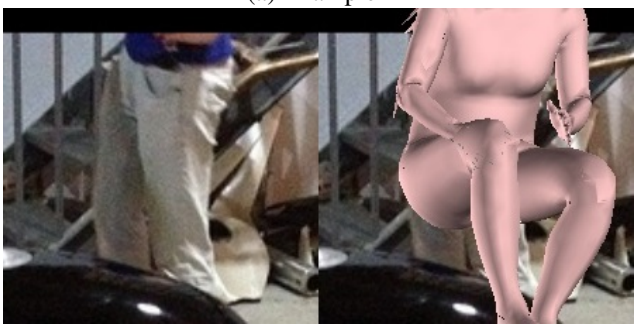
[1] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. In *INTERSPEECH*, 2020. 5

[2] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 4

[3] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 1

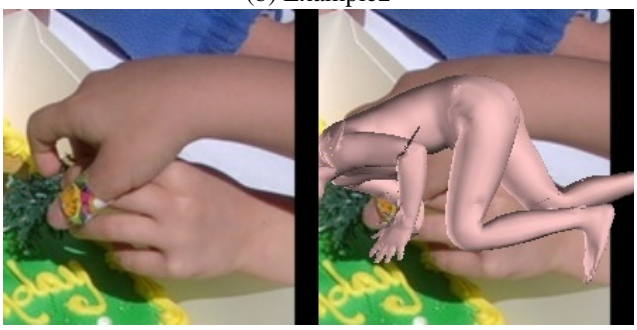
[4] Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. Lite transformer with long-short range attention. In *ICLR*, 2020. 5

(a) Example1



Input Output

(b) Example2



Input Output

Figure 8: Failure cases. Mesh Graphormer may not work well if the reconstruction target is out of the view.