Supplementary Material: MeshMAE: Masked Autoencoders for 3D Mesh Data Analysis

Yaqian Liang[1]*, Shanshan Zhao[2], Baosheng Yu[3], Jing Zhang[3], and Fazhi He[1]

[1] School of Computer Science, Wuhan University, China
[2] JD Explore Academy, China
[3] School of Computer Science, The University of Sydney, Australia
{yqliang, fzhe}@whu.edu.cn, {sshan.zhao00}@gmail.com
{baosheng.yu, jing.zhang1}@sydney.edu.au

In this document, we provide more experimental results, analyses, and visualizations.

1 Backbone Networks

In the main paper, we utilize the ViT-Base as the backbone network. Here, we verify the performance of other Transformer-based backbones, including ViT-Tiny and ViT-Small [1,5]. We train the models from scratch on the ModelNet40 dataset. The classification results in Table 1 indicate that ViT-Tiny and ViT-Small are able to perform well while ViT-Base yields higher scores.

Table 1: Classification results of several Transformer-based backbones.

Model	Layers	Width	MLP	Heads	Acc (%)
ViT-Tiny	12	192	768	3	90.2
ViT-Small	12	384	1536	6	90.8
ViT-Base	12	768	3072	12	91.5

2 Masking Strategy

In the main paper, we mask the patches randomly, while there is another common masking strategy, *i.e.,* block-wise masking. As shown in Figures 1 and 2, block-wise masking removes a very large continuous block. Here, we further investigate the effectiveness of the block-wise masking strategy and make comparison against the random masking under different masking ratios. The results are provided in Table 2. The classification results of random masking are always better than

* This work was done during Y. Liang's internship at JD Explore Academy.

that of block-wise masking, demonstrating that random masking could obtain a better pre-training model, which is consistent with the conclusion in MAE [2]. Besides, we also illustrate the reconstruction results of random masking and block masking in Figures 1 and 2. We can find that the models are able to reconstruct the original shape well for both masking strategies.

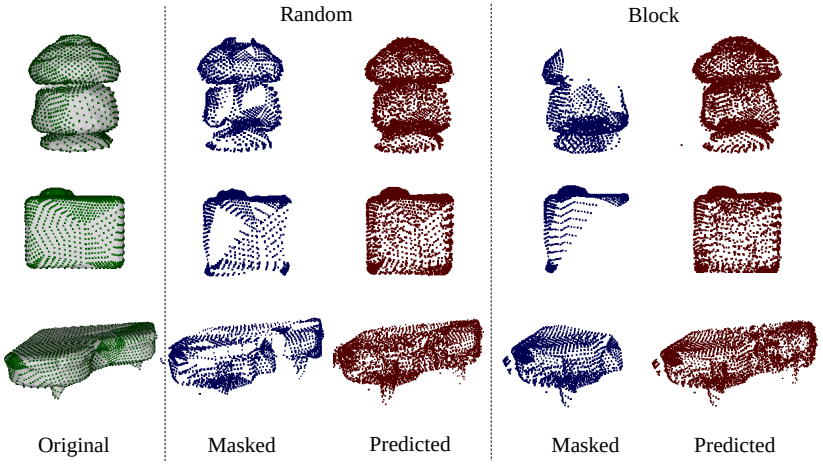

Fig. 1: Reconstruction results of random sampling and block-wise sampling when the masking ratio is 50%.

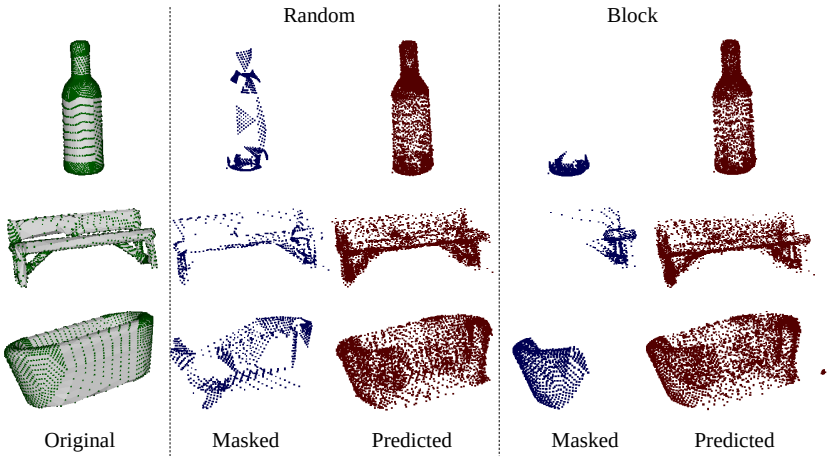

Fig. 2: Reconstruction results of random sampling and block-wise sampling when the masking ratio is 80%.

Table 2: Comparison of different mask strategies, where the pre-training, fine-tuning, and linear probing experiments are all conducted on ModelNet40. 'Fine' indicates the results of fine-tuning, while 'Line' indicates the results of linear probing.

Strategy	Mask	Fine (%)	Line (%)
Random	0.5	91.7	91.1
Block	0.5	91.3	90.6
Random	0.7	91.1	90.6
Block	0.7	90.8	89.2
Random	0.8	91.0	90.2
Block	0.8	90.7	86.3

3 Visualization

3.1 Visualization of Feature Distributions

Here, we visualize the learned features via t-SNE [3] as shown in Figure 3. Figures 3 (a) and (c) show the features obtained by the models pre-trained on ModelNet40 and ShapeNet, respectively. Both of them are visualized on ModelNet40. It is noted that there is no label information during the pre-training process, while the proposed pre-training strategy could already guide the Transformer to learn some semantic information, demonstrating the effectiveness of the proposed pre-training strategy. Then, we finetune these two pre-trained models on ModelNet40, and illustrate the extracted features in Figures 3 (b) and (d), respectively. Through finetuning, the features from different categories are separated better.

3.2 Visualization of Segmentation

In the main paper, we list the comparison of segmentation results quantitatively. Here, we illustrate some visualization examples of the segmentation results in Figure 4. The proposed method could obtain a comparable segmentation performance.

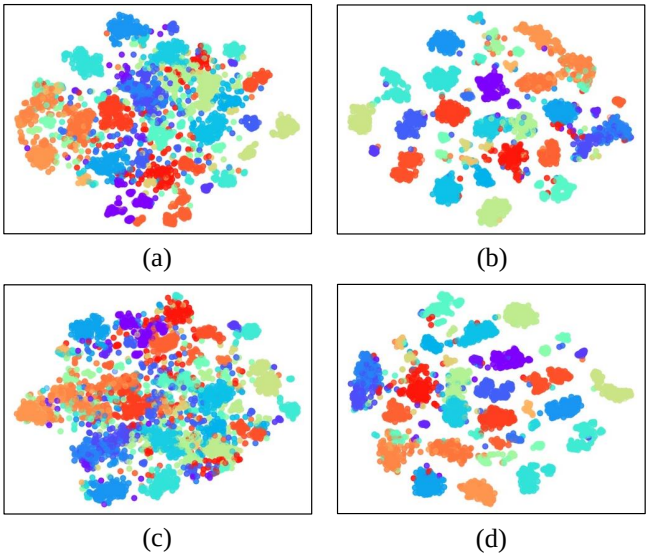

Fig. 3: Visualization of feature distributions. We show the t-SNE visualization of feature vectors learned by MeshMAE. (a) Pre-trained model on ModelNet40; (b) Fine-tuned model (pre-trained on ModelNet40) on ModelNet40; (c) Pre-trained model on ShapeNet; (d) Fine-tuned model (pre-trained on ShapeNet) on ModelNet40. It is noted that different colors indicate different classes.

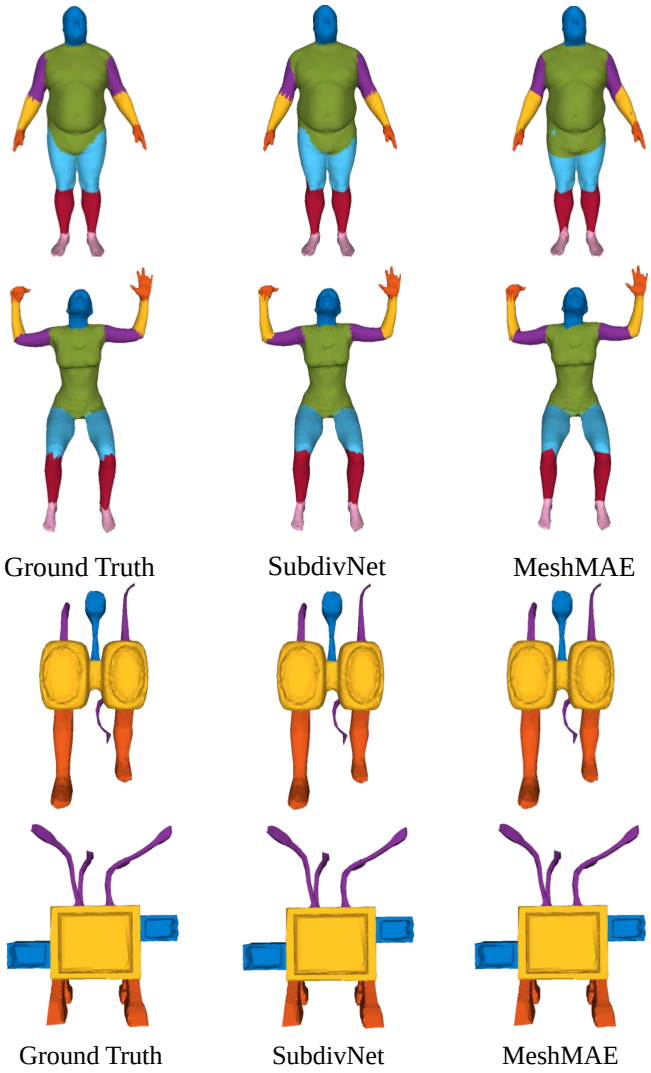

Fig. 4: The comparison of segmentation results. The first two lines are from the Human Body dataset [4], while the last two lines are from the COSEG-aliens dataset [6].

References

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2020)
2. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 16000–16009 (2022)
3. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
4. Maron, H., Galun, M., Aigerman, N., Trope, M., Dym, N., Yumer, E., Kim, V.G., Lipman, Y.: Convolutional neural networks on surfaces via seamless toric covers. ACM Transactions on Graphics (TOG) **36**(4), 71–1 (2017)
5. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: ICLR. pp. 10347–10357 (2021)
6. Wang, Y., Asafi, S., Van Kaick, O., Zhang, H., Cohen-Or, D., Chen, B.: Active co-analysis of a set of shapes. ACM Transactions on Graphics (TOG) **31**(6), 1–10 (2012)