Learning 3D Representations from 2D Pre-trained Models via Image-to-Point Masked Autoencoders Supplementary Material

Renrui Zhang[1,2], Liuhui Wang[3], Yu Qiao[1], Peng Gao[†1], Hongsheng Li[2,4]

[1]Shanghai Artificial Intelligence Laboratory
[2]CUHK MMLab [3]Peking University [4]CPII under InnoHK

{zhangrenrui, gaopeng}@pjlab.org.cn, hsli@ee.cuhk.edu.cn

Dataset	From Scratch	w/o 2D Guidance	**I2P-MAE**
ScanObjectNN [19]	86.34	87.52	**90.11**
ModelNet40 [23]	92.46	93.43	**93.72**
ShapeNetPart [25]	86.38	86.51	**86.76**
ModelNet40-FS [23]	91.20	95.00	**95.50**

Table 1. **Effectiveness of Pre-training.** We report the downstream performance (%) of training from scratch and fine-tuning after pre-training. 'w/o 2D Guidance' denotes the pre-training without learning from 2D pre-trained models. We adopt the PB-T50-RS split of ScanObjectNN and denote the 10-way 20-shot split for few-shot classification as ModeNet-FS.

1. Implementation Details

In this section, we present the detailed model configuration and training settings for fine-tuning I2P-MAE on downstream tasks. All experiments are conducted on a single RTX 3090 GPU.

Shape Classification. For both ModelNet40 [23] and ScanObjectNN [19], we fine-tune I2P-MAE for 300 epochs with a batch size 32. We adopt AdamW [13] optimizer with a learning rate 0.0005 and weight decay 0.05, and utilize cosine scheduler with a 10-epoch warm-up. We append a 3-layer MLP after I2P-MAE's encoder as the classification head. For ScanObjectNN, we adopt max and average pooling to respectively summarize the point tokens from the encoder, and concatenate the two global features along the feature dimension for the head. I2P-MAE takes 2,048 points as input and adopts random scaling with rotation as data augmentation. For ModelNet40, we element-wisely add the two global features for the classification head. I2P-MAE takes 1,024 points as input, and adopts random scaling with translation as data augmentation.

Part Segmentation. On ShapeNetPart [25], we fine-tune I2P-MAE for 300 epochs with a batch size 16. We also adopt AdamW [13] optimizer with a learning rate 0.0002 and weight decay 0.00005, and utilize cosine scheduler with a 10-epoch warm-up. For fair comparison, we experiment with the same segmentation head and training settings as Point-M2AE [27].

2. Additional Ablation Study

Effectiveness of Pre-training. In Table 1, we compare the performance on different downstream tasks between training from scratch and fine-tuning after pre-training. For ScanObjectNN [19], the pure 3D pre-training without image-to-point learning ('w/o 2D Guidance') can improve the classification accuracy by +1.09%, and our proposed 2D-to-3D knowledge transfer further boosts the performance by +2.59%. Similar improvement can be observed on other downstream datasets, which demonstrates the significance of the pre-training of I2P-MAE.

Pre-training with Limited 3D Data. In Figure 1, we compare the performance of pre-training with deficient 3D training data by curves, whose quantitive results are report in Table 5 of the main paper. Guided by 2D pre-trained models, I2P-MAE can acheive comparable performance to Point-M2AE [27] with only half of the 3D data.

Learning Curves of Pre-training. In Figure 3 and 2, we show the comparison of training I2P-MAE from scratch and fine-tuning after pre-training on two shape classification datasets. Our image-to-point pre-training can largely accelerate the convergence speed during fine-tuning and the final classification accuracy, indicating the effectiveness of the 2D-to-3D knowledge transfer.

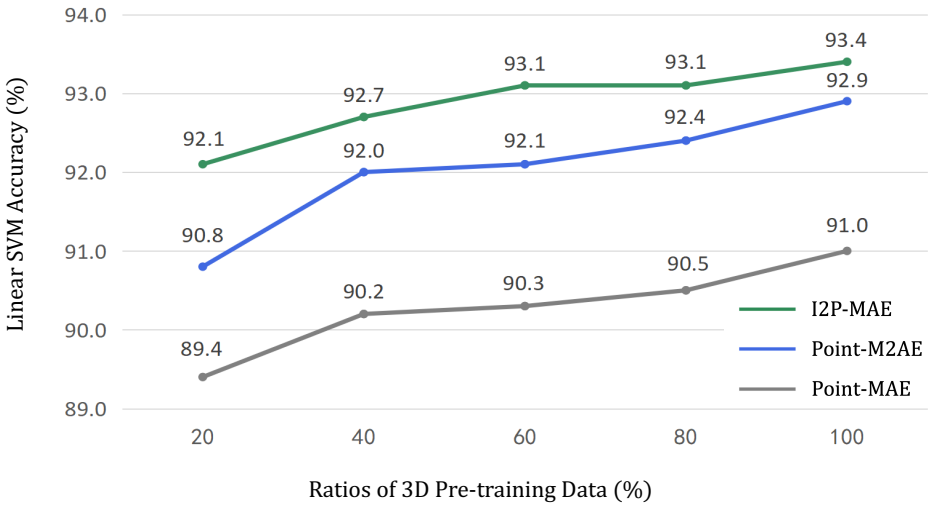

Figure 1. **Pre-training with Limited 3D Data.** We randomly sample different ratios of 3D data in ShapeNet [4] for pre-training and report the linear SVM accuracy on ModelNet40 [23]. I2P-MAE effectively alleviates the need for large-scale 3D data by 2D guidance.

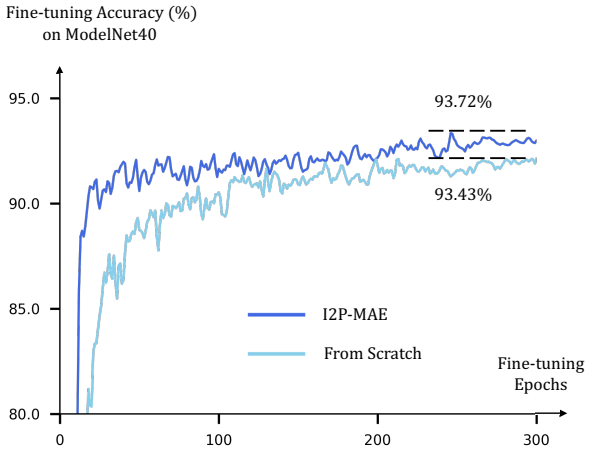

Figure 2. **I2P-MAE Fine-tuning vs. Training from Scratch on ModelNet40 [23].** We report the fine-tuning accuracy on ModelNet40.

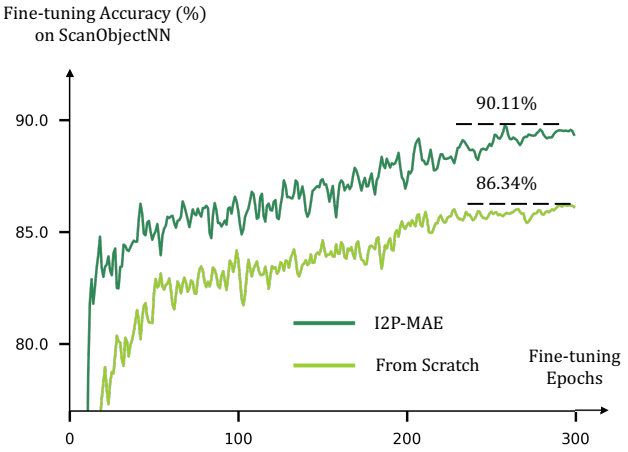

Figure 3. **I2P-MAE Fine-tuning vs. Training from Scratch on SacnObjectNN [19].** We report the fine-tuning accuracy on the PB-T50-RS split of ScanObjectNN.

2D Pre-trained Models. In Figure 4, we compare the guidance performance of 2D models with different architectures and pre-trained methods on the OBJ-BG split of ScanObjectNN. As shown, ViT [6] pre-trained by CLIP [18] performs the best, which learns from 400 million image-text pairs with more sufficient open-world semantics, compared to the pre-training on the closed-set ImageNet-1K [3, 24]. As the transformer [20] is expert at exploring long-range dependencies, ViT and Swin [16] can capture better global spatial cues from the multi-view depth maps for image-to-point learning than ResNet [9].

Projected View Number. In Table 2 (1^{st} and 2^{nd} rows), we show how the number of projected views affect the performance of I2P-MAE. As default, we project the point cloud into 3 views along the x, y, z axes. For the view number 1 and 2, we enumerate all possible projected views along x, y, z axes, and report the highest results in the table. As shown, using less views would harm the pre-training performance, which constrains 2D pre-trained models from 'seeing' complete 3D shapes due to occlusion. Instead, the 3D network can learn more comprehensive high-level semantics from the 2D representations of all three views.

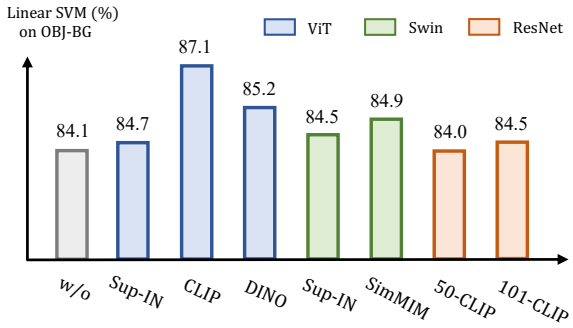

Figure 4. **Different 2D Pre-trained Models.** 'Sup-IN' denotes the supervised pre-training on ImageNet-1K [5]. For ViT [6] and Swin [16], we adopt the base-size models. For ResNet [9], we utilize the 50-layer and 101-layer variants.

3D Attention Cloud and 2D-semantic Target. We first investigate how to aggregate multi-view 2D attention maps as the attention cloud for 2D-guided masking in Table 2 (3^{rd} and 4^{th} rows). Compared to assigning the maximum or minimum score to a certain point, averaging the 2D attention scores from different views achieves the best performance. Then, we explore how to generate the 2D-semantic targets from multi-view 2D features in Table 2 (5^{th} row). The results indicate that, concatenating 2D features between different views performs better than averaging them, which preserves more diverse 2D semantics for reconstruction.

Fine-tuning Settings. In Table 3, we experiment different fine-tuning settings for downstream shape classification on the two datasets. For the point tokens from the encoder, 'Max Only' and 'Ave Only' denote applying either max or average pooling to summarize global features for the classification head. 'Add' or 'Concat' denotes to add or concatenate the two global features after max and average pooling. We observe that, 'Add' and 'Concat' perform the best for ModelNet40 [23] and ScanObjectNN [19], respectively.

3. Additional Related Work

Masked Autoencoders. To achieve more efficient masked image modeling [1, 2, 24, 28], MAE [8] is firstly proposed on 2D images with an asymmetric encoder-decoder transformer [6]. The encoder takes as input a randomly masked image and is responsible for extracting its high-level latent representation. Then, the lightweight decoder explores informative cues from the encoded visible features, and reconstructs raw RGB pixels of the masked patches. Given its superior performance on downstream tasks, a series of follow-up works have been developed to improve MAE with customized designs: pyramid architectures with convolution stages [7], window attention by grouping visible tokens [11], high-level targets with semantic-aware sampling [10], and others [15]. Following the spirit, Point-MAE [17] and MAE3D [12] extend MAE-style pre-training on 3D point clouds, which randomly sample visible point tokens for the encoder and reconstruct masked 3D coordinates via the decoder. Point-M2AE [27] further modifies the transformer architecture to be hierarchical for multi-scale 3D learning. Our proposed I2P-MAE aims to endow masked autoencoding on point clouds with the guidance from 2D pre-trained knowledge. By introducing the 2D-guided masking and 2D-semantic reconstruction, I2P-MAE fully releases the potential of MAE paradigm for 3D representation learning.

Projected Views	3D Attention Cloud	2D-semantic Target	ModelNet40
2	Ave	Concat	93.0
1	Ave	Concat	92.9
3	Max	Concat	93.3
3	Min	Concat	92.8
3	Ave	Ave	93.1
3	Ave	Concat	**93.4**

Table 2. **Different Image-to-Point Settings.** 'Ave', 'Max', 'Min', 'Concat' denote different operations to aggregate multi-view 2D representations. We report the linear SVM accuracy (%).

Settings	ModelNet40 [23]	ScanObjectNN [19]
Max Only	93.31	89.90
Ave Only	93.56	88.83
Add	**93.72**	89.20
Concat	93.23	**90.11**

Table 3. **Fine-tuning Settings.** We experiment different approaches to summarize global features of the encoder for downstream fine-tuning. We report the fine-tuning accuracy (%) on ModelNet40 and the PB-T50-RS split of ScanObjectNN.

4. Few-shot Classification

We fine-tune I2P-MAE for few-shot classification on ModelNet40 [23] in Table 4. Following previous work [17, 26, 27], we adopt the same training settings and few-shot dataset splits, i.e., 5-way 10-shot, 5-way 20-shot, 10-way 10-shot, and 10-way 20-shot. With limited downstream fine-tuning data, our I2P-MAE exhibits competitive performance among existing methods, e.g., +0.5% classification accuracy to Point-M2AE [27] on the 10-way 20-shot split.

Method	5-way		10-way	
	10-shot	20-shot	10-shot	20-shot
DGCNN [22]	91.8 ± 3.7	93.4 ± 3.2	86.3 ± 6.2	90.9 ± 5.1
[P] DGCNN + OcCo [21]	91.9 ± 3.3	93.9 ± 3.1	86.4 ± 5.4	91.3 ± 4.6
Transformer [26]	87.8 ± 5.2	93.3 ± 4.3	84.6 ± 5.5	89.4 ± 6.3
[P] Point-BERT [26]	94.6 ± 3.1	96.3 ± 2.7	91.0 ± 5.4	92.7 ± 5.1
[P] MaskPoint [14]	95.0 ± 3.7	97.2 ± 1.7	91.4 ± 4.0	93.4 ± 3.5
[P] Point-MAE [17]	96.3 ± 2.5	97.8 ± 1.8	92.6 ± 4.1	95.0 ± 3.0
[P] Point-M2AE [27]	96.8 ± 1.8	98.3 ± 1.4	92.3 ± 4.5	95.0 ± 3.0
[P] I2P-MAE	**97.0 ± 1.8**	**98.3 ± 1.3**	**92.6 ± 5.0**	**95.5 ± 3.0**

Table 4. **Few-shot Classification on ModelNet40 [23].** We report the average classification accuracy (%) with the standard deviation (%) of 10 independent experiments. [P] denotes to fine-tune the models after self-supervised pre-training.

Method	1-epoch Time	GPU Mem.	Converge Time
Point-M2AE [27]	11 min	20,339 MiB	43.3 h
I2P-MAE	13 min	22,451 MiB	25.1 h

Table 5. **Pre-training Efficiency Comparison.** We test with a batch size 64 on one RTX 3090 GPU.

5. Discussion

Other Ways using 2D Features? Besides serving as reconstruction target, the multi-view features from 2D pre-trained models can also be utilized to assist the 3D-coordinate reconstruction. To verify this, we back-project the multi-view 2D features into 3D space, and integrate them with corresponding point tokens before feeding into the decoder. In this way, point tokens can leverage sufficient 2D semantics to reconstruct 3D coordinates. The experiments indicate that, by element-wise addition and concatenation with point tokens, the 2D features can boost the 3D learning over Point-M2AE [27] by +0.2% and +0.4%, respectively, for linear SVM on ModelNet40 [23]. Such approach is helpful but inferior to 2D-semantic reconstruction in I2P-MAE with +0.5% performance gain.

Training Resources by Running 2D Models? I2P-MAE requires to concurrently run a pre-trained 2D model, which introduces extra pre-training resources. In Table 5, we compare the pre-training efficiency of I2P-MAE and Point-M2AE. Although I2P-MAE consumes more 1-epoch time and GPU memory, our method converges much faster and takes less overall pre-training time.

I2P-MAE with CLIP-50 Performs Worse than Point-M2AE? In Figure 4, I2P-MAE guided by CLIP's [18] ResNet-50 [9] backbone achieves lower classification accuracy than the baseline Point-M2AE, which is resulted from the following two aspects. **1)** Intuitively, weaker 2D pre-trained models would bring less improvements to I2P-MAE, such as ResNet-50's 84.0% compared to ViT's [6] 87.1% in Figure 4. **2)** To achieve higher pre-training efficiency, we directly adopt projected depth maps as the input for 2D models pre-trained by natural images. Such semantic gap between natural and depth images might disturb the multi-view 2D features and attention maps, especially the weak CLIP-50. Thus, utilizing paired natural 2D images or rendered depth maps for pre-training is expected to improve the performance. We lease this as a future work.

Can 2D Pre-trained Models be Applied to Depth Maps? The 2D models are normally pre-trained by ImageNet [5] containing natural images, such as DINO [3] and SimMIM [24] in Figure 4. For CLIP pre-trained by larger-scale image-text pairs, its training data actually contains a number of depth maps from the Internet, so it provides better guidance for I2P-MAE than other 2D models. To further verify this, we sample a sub-set containing different categories from ShapeNet [4] and directly apply CLIP's ViT backbone for zero-shot classification on the depth maps. It achieves more than 50% accuracy, indicating 2D the pre-trained CLIP can recognize depth maps without training.

6. Additional Visualization

In Figure 5, we additionally visualize the input point cloud, random masking, spatial attention cloud, 2D-guided masking, and the reconstructed 3D coordinates, respectively. As shown, the 2D-guided masking can preserve the semantically important 3D geometries guided by the spatial attention cloud (darker points indicate higher scores). In this way, I2P-MAE can inherit more significant 2D knowledge through the 2D-semantic reconstruction.

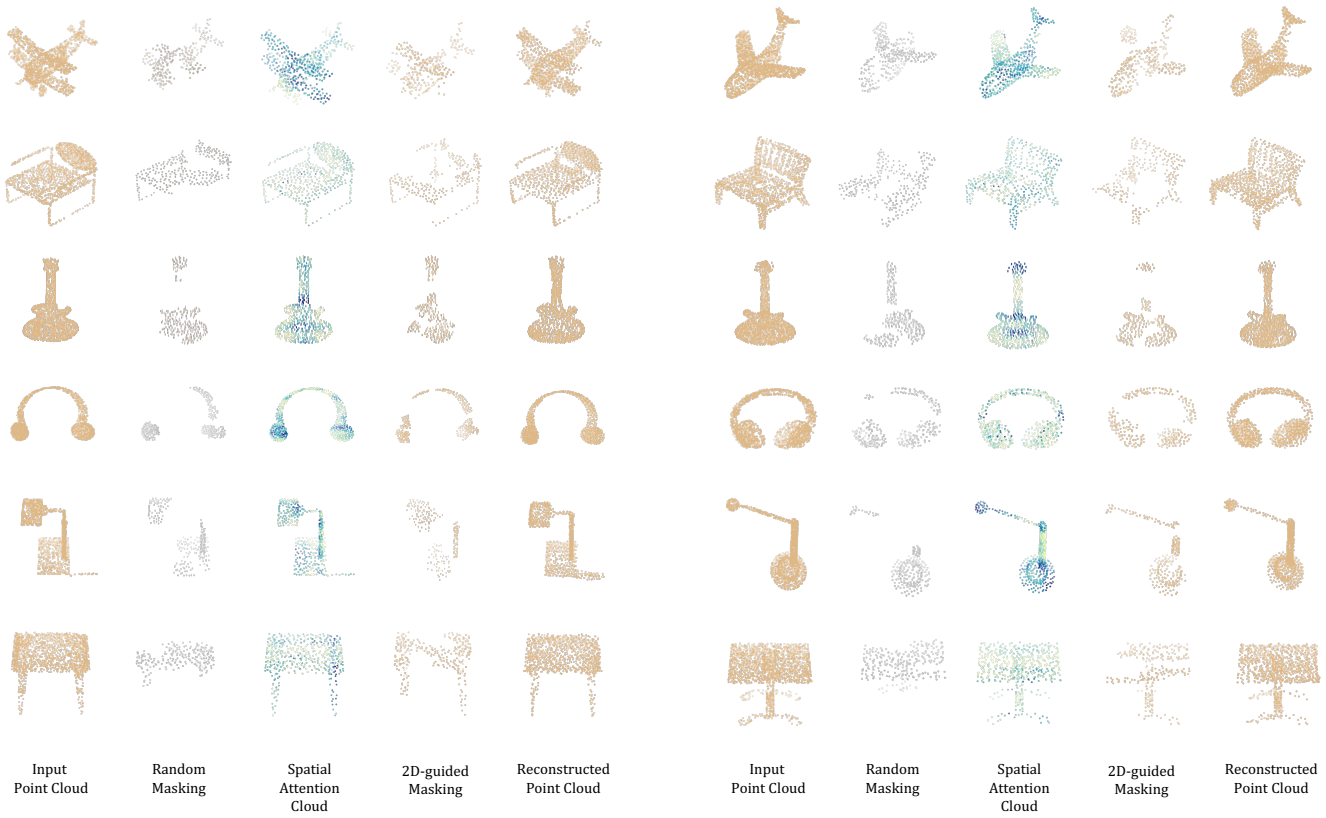

Figure 5. **Additional Visualization of I2P-MAE.** Guided by the spatial attention cloud, I2P-MAE's masking (the 4^{th} and 9^{th} columns) preserves more semantically important 3D structures than random masking.

References

[1] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022. 3

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2, 4

[4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2, 4

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3, 4

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3, 4

[7] Peng Gao, Teli Ma, Hongsheng Li, Jifeng Dai, and Yu Qiao. Convmae: Masked convolution meets masked autoencoders. *arXiv preprint arXiv:2205.03892*, 2022. 3

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 3

[9] 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*, pages 770–778, 2016. 2, 3, 4

[10] Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and Sun-Yuan Kung. Milan: Masked image pretraining on language assisted representation. *arXiv preprint arXiv:2208.06049*, 2022. 3

[11] Lang Huang, Shan You, Mingkai Zheng, Fei Wang, Chen Qian, and Toshihiko Yamasaki. Green hierarchical vision transformer for masked image modeling. *arXiv preprint arXiv:2205.13515*, 2022. 3

[12] Jincen Jiang, Xuequan Lu, Lizhi Zhao, Richard Dazeley, and Meili Wang. Masked autoencoders in 3d point cloud representation learning. *arXiv preprint arXiv:2207.01545*, 2022. 3

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[14] Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. *arXiv preprint arXiv:2203.11183*, 2022. 4

[15] Jihao Liu, Xin Huang, Yu Liu, and Hongsheng Li. Mixmim: Mixed and masked image modeling for efficient visual representation learning. *arXiv preprint arXiv:2205.13137*, 2022. 3

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 3

[17] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. *arXiv preprint arXiv:2203.06604*, 2022. 3, 4

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 4

[19] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1588–1597, 2019. 1, 2, 3

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[21] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J Kusner. Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9782–9792, 2021. 4

[22] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019. 4

[23] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 1, 2, 3, 4

[24] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. *arXiv preprint arXiv:2111.09886*, 2021. 2, 3, 4

[25] Li Yi, Vladimir G Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016. 1

[26] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. *arXiv preprint arXiv:2111.14819*, 2021. 3, 4

[27] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: Multi-scale masked autoencoders for hierarchical point cloud pre-training. *arXiv preprint arXiv:2205.14401*, 2022. 1, 3, 4

[28] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 3