Point Cloud Self-supervised Learning via 3D to Multi-view Masked Learner

Zhimin Chen[1] Xuewei Chen[1] Xiao Guo[2] Yingwei Li[3]
Longlong Jing[4] Liang Yang[4] Bing Li[1]
[1]Clemson University [2]Michigan State University [3]Johns Hopkins University [4]The City University of New York

1. Experiments

Settings. We utilize the ShapeNet [1] for pre-training. To obtain a dense depth map, the input point number N is set as the baseline method. The number of the projection view K is set to 3 and the depth map size is set as 224×224. Random scaling and random rotation are implemented as data augmentation during pre-training. We project point clouds into multi-view after the augmentation. Our method employs an AdamW optimizer [9] and cosine learning rate decay [8]. The network is trained for 300 epochs with a batch size of 128. The initial learning rate, weight decay, and mask ratio are set to 2×10^{-4}, 0.05, and 0.7. The scale for 3D and 2D modality is $[2, 4, 8, 16, 32, 64]$ and $[6, 12, 24, 49, 98, 196]$.

Few-shot Learning. We conducted few-shot learning experiments on the ModelNet40 dataset [18] using the n-way, m-shot setting, following the protocol of Point-MAE. During training, we randomly selected n classes and m objects from each class. During testing, we randomly selected 20 unseen objects from each of the n classes for evaluation. We conducted 10 independent experiments for each setting and reported the mean accuracy with standard deviation. The results of our fine-tuned few-shot classification are shown in Table 1. Our method outperformed the baseline and state-of-the-art methods in all settings.

Linear SVM Result To evaluate the transfer capacity, we directly utilize the features extracted by I2P-MAE's encoder for linear SVM on the synthetic ModelNet40 [18] without any fine-tuning or voting. The results on ModelNet40 are shown in Table 2. It shows that our RECON outperforms the last SOTA method I2P-MAE [22] by 0.3% even without using pre-trained foundation models. This improvement in SVM classification performance underscores the efficacy of our approach in learning superior quality 3D representations and highlights the value of the inherent multi-view property of 3D data.

2. Methods Comparison

Comparison with I2P-MAE [22], ReCon [11], and ShapeLLM [12]. I2P-MAE, ReCon, and ShapeLLM focus primarily on two types of features: 3D geometric features and semantic/textual representations. They employ MAE-based structures to reconstruct the original point clouds, thereby capturing detailed 3D geometric data. Additionally, they utilize techniques such as contrastive learning or knowledge distillation to extract semantic and textual features from 2D images and language models. These methods directly adopt existing 3D MAE frameworks specifically, I2P-MAE utilizes Point-M2AE, while ReCon and ShapeLLM leverage Point-MAE for geometric representation and their innovation lies in the novel use of foundation models for knowledge distillation.

I2P-MAE performs pixel-to-3D token knowledge distillation by adding additional layers after the M2AE encoder, calculating MSE loss between the point tokens and 2D pixel-level features derived from foundation models. ReCon uses Point-MAE as the base structure to reconstruct original point clouds from masked point cloud inputs, while also incorporating instance-level contrastive learning to distill knowledge from both text and image foundation models. ShapeLLM builds upon ReCon by using larger models with more parameters, leveraging large language models to enable advanced 3D reasoning. **In contrast**, our approach focuses on advancing geometric learning in 3D self-supervised learning (SSL), *emphasizing the use of the inherent multi-view attributes in point cloud data to enhance geometric understanding*, solely within the 3D modality. Due to the fundamental differences in goals and methodologies, a direct comparison with I2P-MAE and ReCon would not provide a fair evaluation.

Comparison with 3D Geometric Learning SSL Methods. Our work focuses on 3D geometric learning without leveraging foundation models, similar to methods like Point-M2AE [21], Point-GPT [3], Pi-MAE [2], Joint-MAE [6], and TAP [15], which aim to learn pure 3D geometric representations without relying on knowledge distillation from foundation models. Existing MAE-based 3D geometric learning methods generally follow two modification directions: (1) Encoder structure modification, as seen in methods like Point-M2AE and Point-GPT, and (2) Incorporating 2D information into the reconstruction process, as

Method	Foundation Model Needed	5-way		10-way	
		10-shot	20-shot	10-shot	20-shot
Point-BERT [20]	×	94.6 ± 3.1	96.3 ± 2.7	91.0 ± 5.4	92.7 ± 5.1
Point-MAE [10]	×	96.3 ± 2.5	97.8 ± 1.8	92.6 ± 4.1	95.0 ± 3.0
Joint-MAE [6]	×	96.7 ± 2.2	97.9 ± 1.8	92.6 ± 3.7	95.1 ± 2.6
Point-M2AE [21]	×	96.8 ± 1.8	98.3 ± 1.4	92.6 ± 5.0	95.0 ± 3.0
TAP [15]	×	97.3 ± 1.8	97.8 ± 1.7	93.1 ± 2.6	95.8 ± 1.0
Ours(Point-MAE)	×	97.3 ± 1.9	98.2 ± 1.6	93.2 ± 4.1	96.0 ± 2.7
Ours (Point-M2AE)	×	**97.6 ± 2.1**	**98.5 ± 1.3**	**93.6 ± 3.9**	**96.1 ± 2.1**
ACT [5]	✓	96.8 ± 2.3	98.0 ± 1.4	93.3 ± 4.0	95.6 ± 2.8
I2P-MAE [22]	✓	97.0 ± 1.8	98.3 ± 1.4	92.3 ± 4.5	95.5 ± 3.0
ReCon [11]	✓	**97.3 ± 1.9**	**98.9 ± 1.2**	**93.3 ± 3.9**	**95.8 ± 3.0**

Table 1. Few-shot classification performance on ModelNet40 [18], measured by the accuracy (%) and standard deviation (%). $*$ denotes the model without pre-training. [Key: **Best results**, Second best results.].

Method	ModelNet40
Transformer + OcCo [20]	89.6
Point-BERT [20]	87.4
Point-MAE [10]	91.0
Joint-MAE [6]	92.4
Point-M2AE [21]	92.9
I2P-MAE [22]	93.4
Ours + Point-MAE	93.1
Ours + Point-M2AE	**93.7**

Table 2. Linear SVM Classification on ModelNet40 [18]. We compare the accuracy (%) of existing self-supervised methods.

done by Pi-MAE, Joint-MAE, and TAP.

Our work follows the second direction but addresses significant limitations in existing methods that leverage 2D information for 3D geometric learning. Specifically, approaches like Pi-MAE, Joint-MAE, and TAP do not fully exploit the multi-view properties of 3D point clouds and their inherently multi-modal attributes. For example, a point cloud can be directly projected into multi-view images using pose information. Incorporating masked 2D images as input, as done by Pi-MAE and Joint-MAE during the MAE training process, is unnecessary and potentially detrimental, as it can cause the network to over rely on visible 2D information to predict masked content rather than developing a comprehensive understanding of multi-view geometry, ultimately degrading the quality of learned 3D representations. Moreover, TAP uses a pretrained VAE to reconstruct 2D images from 3D inputs but fails to effectively leverage multi-view information. In contrast, our method

Method	OBJ-BG	OBJ-ONLY	PB-T50-RS
Stage 1 + MAE	92.34	91.88	87.56
Ours	**93.32**	**92.69**	**88.93**

Table 3. Ablation study for the two-stage design.

introduces a unified approach that uses masked point clouds to reconstruct both multi-view 2D images and the original point clouds, ensuring a more comprehensive understanding of 3D geometry while effectively utilizing the multi-view attributes of 3D data. Furthermore, we propose MSMH decoder to better global and local features and a two-stage self-training method to learn well-aligned representations. It is worth to mention that during the fine-tuning and inference stages, we remove additional components, such as the projection layers and MSMH decoder, maintaining the same architecture as Point-MAE to ensure a fair comparison.

3. Additional Ablation Study

The Effectiveness of the Two-stage Framework. As discussed in previous work [4], latent space prediction is highly effective for SSL representation learning. Inspired by this, we introduce latent space prediction into the multi-modality MAE in this paper. Rather than predicting raw inputs across diverse modalities, our model uses masked-view inputs to jointly predict contextualized 2D and 3D representations within a latent space aligned by a teacher model with complete-view inputs. This ensures that student models learn well-aligned and contextualized representations across modalities. Our ablation study, shown in

# Pose Pool Size	PB-T50-RS
3	87.17
6	88.29
12	**88.93**
24	88.54
36	87.76

Table 4. Ablation study for the number of pose pool size on the 3D object classification tasks in ScanObjectNN dataset.

# Recon View Size	PB-T50-RS
1	87.05
2	88.41
3	**88.93**
4	88.12
5	87.58

Table 5. Ablation study for the number of reconstructed views on the 3D object classification task in ScanObjectNN dataset.

Image Type	PB-T50-RS
Depth Image	**88.93**
Rendered Image	88.12

Table 6. Ablation study for the image type on the 3D object classification tasks in ScanObjectNN dataset.

View Configuration	PB-T50-RS
Circular	**88.93**
Spheric	87.97
Spheric & Circular	88.35
Random	87.41

Table 7. Ablation study for the view configuration of the depth images on the 3D object classification tasks in ScanObjectNN dataset.

Table 3, demonstrates the advantage of this two-stage latent prediction method. Compared to the direct reconstruction of masked multi-modality raw inputs in one stage, our approach achieves 1.37 % improvement.

The Effectiveness of Poses Pool Size. The pose pool size represents the total number of poses that can be leveraged in our 3D to multi-view MAE method. The ablation study detailed in Table 4 investigates the impact of varying the number of views in the network on 3D object classification performance, using the ScanObjectNN dataset. The study examines a range of views: 3, 6, 12, 24, and 36 to understand how they affect classification accuracy. The results reveal a notable trend: as the number of views increases, there's generally an improvement in classification accuracy, achieving the best performance at 12 views. Beyond this optimal point, however, the performance decreases with the increase of projected views. This pattern indicates that while increasing the number of views contributes positively to the network's understanding and representation of 3D objects, there is a point beyond which additional views do not yield further benefits. This is because too many views introduce the redundancy of view specific information, leading to a slight decrease in the network's efficiency.

Effectiveness of Image Type. In the ablation study presented in Table 6 , we analyze two commonly used image types for 3D understanding: rendered images and depth images. The results indicate that using depth images yields the best performance, which aligns with findings from previous work, such as Joint-MAE.

The Effectiveness of Network Reconstructed View Numbers. Our method enhances multi-view understanding by randomly selecting several view poses from the pose pool mentioned above, enabling the model to reconstruct corresponding multiple projected depth images. This ablation study focuses on finding the optimal number of reconstructed views for enhancing 3D representation learning in the ScanObjectNN [14] dataset. We examined the impact of the number of reconstructed views from one to five on classification performance in PB-T50-RS setting. According to the results in Table 5, accuracy consistently increases with the number of views, peaking at 3 views. Beyond this point, however, the trend indicates a decrease in performance. This suggests that multiple reconstructed views enhance the network's understanding of multi-view information. However, too many reconstructed views will make the length of the input sequences processed by the decoder very large, thus impacting the network's learning efficiency and capacity.

The Effectiveness of View Configurations. In the ablation study shown in Table 7, different view configurations of depth images for our method in 3D representation learning are analyzed using the ScanObjectNN dataset. The most common view configurations for depth image projection are circular which alignes viewpoints on a circle around the object [13, 19] and spherical which alignes equally spaced viewpoints on a sphere surrounding the object [7, 16]. We test Circular, Spheric, a combination of both, and Random configurations. The Circular configuration proves most effective, achieving the highest accuracies of 88.93 in PB-T50-R, likely due to its comprehensive coverage and consistent viewing angles. The Spheric configuration, while offering a broad perspective, falls slightly short in comparison. Combining Spheric and Circular views improves performance but does not outperform the Circular configuration alone. The Random configuration shows the least effectiveness. This study highlights the Circular view config-

Pose Type	PB-T50-RS
Index	**88.93**
Camera Matrix	88.33

Table 8. Ablation study for the pose type on the 3D object classification tasks in ScanObjectNN dataset.

Masking Ratio	PB-T50-RS
0.6	88.02
0.65	88.46
0.7	**88.93**
0.75	88.15
0.8	87.62

Table 10. Ablation study for the masking ratio on the 3D object classification tasks in ScanObjectNN dataset.

Rec Type	PB-T50-RS
Masked Only	**88.93**
Full	88.41
Visible Only	87.74

Table 9. Ablation study for the feature reconstruction type on the 3D object classification tasks in ScanObjectNN dataset.

Method	**PB-T50-RS**
Point-MAE	85.18
Point-MAE + GVA	85.48
Point-MAE + MSMH	86.03
Ours (with GVA)	88.29
Ours (with MSMH)	**88.93**

Table 11. Comparison of different methods on various evaluation metrics.

uration's superiority in providing a balanced and thorough representation of 3D objects, essential for better representation learning.

The Effectiveness of Pose Type. The ablation study detailed in Table 8 critically examines the influence of pose type on the accuracy of 3D object classification within the ScanObjectNN dataset. It delves into two distinct pose types: Index and Camera Matrix, assessing their effectiveness in PB-T50-R setting of the ScanObjectNN dataset. The Index pose type employs fixed indexes to denote specific pose views, whereas the Camera Matrix approach directly inputs the camera matrix into the pose encoding process to derive pose embeddings. Notably, both pose types demonstrate commendable performance, with the Index slightly surpassing the Camera Matrix. This marginal difference underscores the robustness of the classification method to variations in pose type input, suggesting a flexible adaptability to different pose representation strategies in 3D representation learning.

The Effectiveness of Reconstruction Type. In this research, we leverage the student branch to reconstruct the representations of masked tokens based on guidance from the teacher branch. Our ablation study, presented in Table 9, meticulously evaluates the influence of various reconstruction (Rec) methodologies on the 3D object classification accuracy using the ScanObjectNN dataset. This study differentiates between three reconstruction types: 'Masked Only', 'Full', and 'Visible Only'. The findings indicate that focusing on reconstructing only the masked features yields the most favorable outcomes. In contrast, the approach of reconstructing only the visible features, similar to the previous state-of-the-art method I2P-MAE, results in the least effective performance. These results underscore the effectiveness of our proposed method in more accurately aligning the latent spaces of the teacher and student models and the better ability to fully utilize the multi-view information.

The Effectiveness of Masking Ratio. The ablation study outlined in Table 10 evaluates the effect of different masking ratios on 3D representation learning in the ScanObjectNN dataset. Five masking ratios are tested: 0.6, 0.65, 0.7, 0.75, and 0.8, assessing their impact on performance in PB-T50-RS setting in the ScanObjectNN dataset. The results indicate a clear pattern. As the masking ratio decreases from 0.6 to 0.7, classification accuracy consistently improves. The best performance is observed at a masking ratio of 0.7, with accuracies reaching 88.93%. However, reducing the masking ratio further to 0.8 results in a slight decrease in accuracy. These findings suggest that an optimal masking ratio exists, where a balance is struck between challenging the network sufficiently to learn robust features and retaining enough information for accurate classification. Too much masking may obscure critical details, while too little may not provide enough complexity for effective learning.

Comparison with GVA and MSMH. The motivations and implementations of MSMH and Grouped Vector Attention (GVA) [17] differ fundamentally. While GVA focuses on improving model efficiency and generalization, MSMH is designed to effectively capture both local and global contextual information by organizing distinct, non-overlapping local groups at multiple scales within the reconstructed features. In terms of implementation, GVA divides only the value vector into different groups while applying the same scalar attention weight across them. This design reduces the number of parameters, thereby enhancing efficiency. In contrast, MSMH divides the query, key, and value tokens

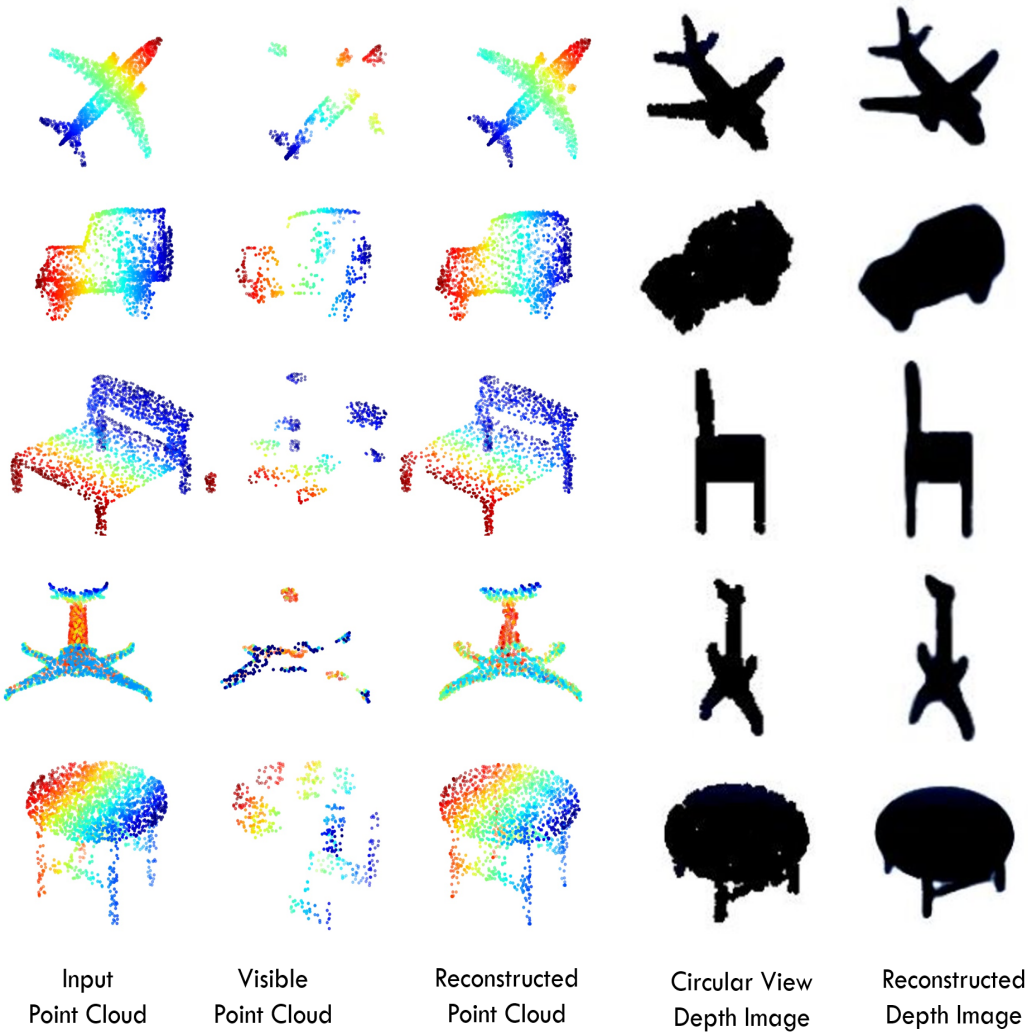

Figure 1. Visualization of 3D to multi-view masked autoencoder (Stage 1 with MAE). Our method not only can reconstruct point clouds from masked input but also generate multi-view depth images.

into distinct, non-overlapping local groups and applies self-attention within each subgroup, rather than across all individual tokens. Additionally, MSMH introduces a multi-scale grouping strategy, where smaller groups capture fine-grained local details, while larger groups capture broader global context. These multi-scale attention features are then concatenated to ensure a more comprehensive understanding of both local and global structures. As a result, MSMH provides a more effective integration of multi-scale spatial relationships than GVA, leading to superior performance. To further validate its effectiveness, we incorporated GVA into our framework for comparison. Experimental results in Table. 11 indicate that while GVA brings only marginal improvement, MSMH achieves significantly better performance, underscoring its advantage in capturing multi-scale geometric structures.

4. Visualization

For the second-stage design, our method focuses on feature reconstruction. Therefore, visualizing the reconstruction across the entire two-stage process poses significant challenges. To address this, we provide visualization results by directly integrating MAE into the stage-one framework, as detailed in Table 7 of the main paper. The visualization results are presented in Fig. 1, where each row illustrates the input point clouds, masked point clouds, reconstructed point clouds, projected depth images, and reconstructed images, respectively. Our method demonstrates the ability to not only reconstruct point clouds from masked inputs but also generate multiview depth images, highlighting its capability to effectively capture the intrinsic multi-modal information of point clouds.

References

[1] 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. 1

[2] Anthony Chen, Kevin Zhang, Renrui Zhang, Zihan Wang, Yuheng Lu, Yandong Guo, and Shanghang Zhang. Pimae: Point cloud and image interactive masked autoencoders for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5291–5301, 2023. 1

[3] Guangyan Chen, Meiling Wang, Yi Yang, Kai Yu, Li Yuan, and Yufeng Yue. Pointgpt: Auto-regressively generative pre-training from point clouds. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[4] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2

[5] Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jianjian Sun, Zheng Ge, Li Yi, and Kaisheng Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? *arXiv preprint arXiv:2212.08320*, 2022. 2

[6] Ziyu Guo, Xianzhi Li, and Pheng Ann Heng. Joint-mae: 2d-3d joint masked autoencoders for 3d point cloud pre-training. *arXiv preprint arXiv:2302.14007*, 2023. 1, 2

[7] Asako Kanezaki, Yasuyuki Matsushita, and Yoshifumi Nishida. Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5010–5019, 2018. 3

[8] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[10] 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. 2

[11] Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. *arXiv preprint arXiv:2302.02318*, 2023. 1, 2

[12] Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *European Conference on Computer Vision*, pages 214–238. Springer, 2025. 1

[13] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 3

[14] 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. 3

[15] Ziyi Wang, Xumin Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. Take-a-photo: 3d-to-2d generative pre-training of point cloud models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5640–5650, 2023. 1, 2

[16] Xin Wei, Ruixuan Yu, and Jian Sun. View-gcn: View-based graph convolutional network for 3d shape analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1850–1859, 2020. 3

[17] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35:33330–33342, 2022. 4

[18] 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

[19] Tan Yu, Jingjing Meng, and Junsong Yuan. Multi-view harmonized bilinear network for 3d object recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 186–194, 2018. 3

[20] 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. 2

[21] 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, 2

[22] Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21769–21780, 2023. 1, 2