Supplementary Material

1. Comparisons to more peer methods

Apart from the self-reconstruction methods in the main paper (including baseline Point-MAE [14] and others), there are some other peer methods, including Point-FEMAE [22], PCP-MAE [27], I2P-MAE [24], Joint-MAE [6], Cross-BERT [11], and TAP [19]. Here, we discuss the relation of Point-PQAE with these approaches and compare its performance against them to better position our work. A brief comparison of peer methods with our Point-PQAE can be seen in Tab. 1, and the performance of these methods on downstream tasks is reported in Tab. 2. Point-PQAE achieves the best or comparable performance when compared with them.

Relation to Point-FEMAE [22]. Connection. Both of them are reconstruction-based methods. **Differences. 1)** *Pre-Training Efficiency.* Point-FEMAE performs mask reconstruction in both the global and local branches and introduces Local Enhancement Module (LEM) which consists of some convolution layers and MLP layers to each transformer block. To achieve local patch convolution with coordinate-based nearest neighbors, when tokens are input to LEM, it duplicates K nearest neighboring patches ($k = 20$) for each input token and aggregates nearest information for each token which brings an extra calculation burden to each block in the encoder. Point-PQAE utilizes original transformer blocks, making it more efficient during pre-training. **2)** *Backbone.* Point-FEMAE reserves the LEM modules in the encoder for fine-tuning which means adding convolution and MLP layers for each transformer block to the backbone for fine-tuning, while our Point-PQAE utilizes an encoder consisting of pure transformer blocks for fine-tuning, which remains simple and is aligned to previous work.

Relation to PCP-MAE [27]. Connection. Both of them are reconstruction-based methods. **Differences.** Targeted at alleviating information leakage of centers in point cloud, PCP-MAE proposes a new module called Predicting Center Module (PCM) and a novel loss for better utility of centers based on Point-MAE, which is still a self-reconstruction method. Our Point-PQAE differs from it as it is a pioneering cross-reconstruction method; it uses VRPE to perform cross-view point cloud reconstruction, which overcomes the limitations of self-reconstruction methods.

Relation to I2P-MAE [24]. Connection. Both of them are reconstruction-based methods. **Differences.** I2P-MAE heavily relies on strong pre-trained 2D models as a guide to achieve multi-task cross-modal learning while our Point-PQAE utilizes single-modal data without relying on any pre-trained model. Besides, the utilization of 2D data in I2P-MAE brings a heavy computation burden to the pre-training process.

Relation to Joint-MAE [6]. Connection. Both of them are reconstruction-based methods. **Differences.** Joint-MAE utilizes a shared weight encoder but 3 different decoders for pre-training, and it's a multi-task pre-training method including 2D / 3D / 2D-3D reconstruction, which makes it more computational while Point-PQAE utilizes one encoder and one decoder only for per-training and is a single task method which focuses on 3D data cross-reconstruction.

Relation to Cross-BERT [11]. Connection. Both are point cloud self-supervised methods. **Differences.** Cross-BERT is a method that utilizes two isolated encoders for cross-modal learning of point clouds and rendered images. To prevent the collapse of its intra-/cross- modal contrastive learning, it further uses another two momentum encoders that perform EMA-update, which makes pre-training of Cross-BERT much more complex than Point-PQAE. Additionally, Cross-BERT requires pre-training a dVAE as the tokenizer and includes a mask cross-modal learning task alongside contrastive learning. In contrast, Point-PQAE focuses solely on point clouds, utilizing a single encoder for single-task self-supervised pre-training.

Relation to TAP [19]. Connection. Both are reconstruction-based methods. **Differences.** TAP makes cross-modal reconstruction, which renders images of point clouds with different poses and after getting the latent representation of the 3D point cloud, it uses the pose information of the rendered images to query cross-modal information from the latent representation by cross-attention. And then using the queried information to rebuild the rendered image. Point-PQAE uses view-relative position embedding (VRPE) to make cross-view information interaction by cross-attention to achieve cross-reconstruction which uses 3D data only. Extra online operation on rendering and processing 2D data in TAP will raise the computational needs compared to Point-PQAE.

Table 1. Methodology comparisons between our Point-PQAE and other peer methods. The "Extra" in the table means extra parts in contrast to Point-MAE which adopts standard transformer blocks as backbone.

Methods	Single-/Cross-Modal	Pre-trained Model Needed	Single-/Multi-Task	Extra Transformer Blocks (Pre-training)	Extra Modules (Fine-tuning)
Point-MAE [14]	Single	✗	Single	✗	✗
Point-FEMAE [22]	Single	✗	Multi	✗	✓
PCP-MAE [27]	Single	✗	Multi	✗	✗
I2P-MAE [24]	Cross	✓	Multi	✗	✗
Joint-MAE [6]	Cross	✗	Multi	✓	✗
Cross-BERT [11]	Cross	✓	Multi	✓	✗
TAP [19]	Cross	✗	Single	✗	✗
Point-PQAE	Single	✗	Single	✗	✗

Table 2. Performance of peer methods. The classification results on ScanObjectNN and ModelNet40 and few-shot learning results on ModelNet40 are reported by accuracy (%). We term OBJ_BG, OBJ_ONLY, PB_T50_RS as BG, OY, RS respectively. We compare methods using the • plain Transformer architectures, *e.g.* Point-MAE[14], Point-PQAE (ours), ○ hierarchical Transformer architectures and ○ methods with extra modules during fine-tuning.

Methods	#P	ScanObjectNN			ModelNet40		ModelNet40 few-shot			
							5-way		10-way	
		BG	OY	RS	1K P	8K P	10-shot	20-shot	10-shot	20-shot
•Point-MAE [14]	22.1	90.0	88.3	85.2	93.8	94.0	96.3±2.5	97.8±1.8	92.6±4.1	95.0±3.0
○Point-FEMAE [22]	27.4	95.2	93.3	90.2	**94.5**	-	97.2±1.9	98.6±1.3	**94.0±3.3**	95.8±2.8
•PCP-MAE [27]	22.1	**95.5**	**94.3**	**90.4**	94.2	-	**97.4±2.3**	**99.1±0.8**	93.5±3.7	95.9±2.7
○I2P-MAE [24]	-	94.2	91.6	90.1	94.1	-	97.0±1.8	98.3±1.3	92.6±5.0	95.5±3.0
○Joint-MAE [6]	-	90.9	88.9	86.1	94.0	-	96.7±2.2	97.7±1.8	92.6±3.7	95.1±2.6
•Cross-BERT [11]	22.1	93.7	92.1	89.0	94.2	**94.4**	97.0±2.1	98.2±1.3	93.0±3.4	95.6±3.0
•TAP [19]	22.1	90.4	89.5	85.7	-	-	97.3±1.8	97.8±1.7	93.1±2.6	95.8±1.0
•Point-PQAE	22.1	95.0	93.6	89.6	93.9	94.3	96.9±3.0	99.0±1.0	**94.0±4.0**	**96.1±2.8**

2. Related cross-reconstruction works

Our Point-PQAE pioneers the cross-reconstruction paradigm in 3D point cloud self-supervised learning (SSL). **There are two essential components in our proposed cross-reconstruction framework:** 1) Two isolated/decoupled views, rather than two parts of the same instance that maintain a fixed relative relationship. 2) A model that achieves cross-view reconstruction using relative position information. To our knowledge, there are no similar previous methods in this domain. To better position our methodology, we compare it to similar SSL methods, including SiamMAE [7] and CropMAE [4], proposed in the image domain. SiamMAE operates on pairs of randomly sampled video frames and asymmetrically masks them, utilizing the past frame to predict the masked future frame. CropMAE relies on image augmentations to generate two views, using one to reconstruct the other.

Relation of SiamMAE and CropMAE to our Point-PQAE: All are cross-reconstruction methods. Both SiamMAE and CropMAE have two essential components for cross-reconstruction-framework including two-view (different frames sampled from one video for SiamMAE and isolated augmented images for CropMAE) and cross-view reconstruction model. They can be treated as cross-reconstruction methods, similar to our Point-PQAE.

Difference of SiamMAE and CropMAE to our Point-PQAE: 1) Different domain: SiamMAE and CropMAE focus on the 2D SSL domain. Our Point-PQAE is the first method for cross-reconstruction in the 3D SSL domain. **2) Asymmetric/symmetric reconstruction:** SiamMAE uses the past to predict the future, which is asymmetric. CropMAE performs asymmetric reconstruction and doesn't explore siamese cross-reconstruction. In contrast, our Point-PQAE is inherently symmetric, and the siamese loss brings a performance gain. **3) No relative information utilized:** SiamMAE and CropMAE do not incorporate relative information into training but rely on non-fully masking to guide the cross-reconstruction. The VRPE adopted by our Point-PQAE provides explicit guidance, making training more stable and improving explainability. **4) No tuned-needed mask ratio exists in our framework.** There is a hyperparameter mask ratio that needs to be tuned in both SiamMAE and CropMAE, but this is not the case in our Point-PQAE framework.

Relation and differences between Joint-MAE [6] and PiMAE [2]. We discuss the differences between our framework and two seemingly similar methods in the point cloud domain: Joint-MAE [6] and PiMAE [2]. Joint-MAE and PiMAE adopt a similar strategy, utilizing paired point clouds and images to perform cross-modal masked autoencoding. Our framework, however, differs significantly from these two methods.

The **relation** of these methods to our Point-PQAE is that all three are self-supervised approaches that focus on the point cloud domain.

Differences:

1) Different motivations: Joint-MAE and PiMAE aim to explore the semantic correlation between 2D and 3D data by performing 3D-2D interactions and achieving cross-modal self-reconstruction through cross-modal knowledge. In contrast, inspired by the success of two-view pre-training paradigms, we propose Point-PQAE, the first cross-reconstruction framework for point cloud self-supervised learning (SSL).

2) Different modalities: Both Joint-MAE and PiMAE rely on paired image-point cloud data, making them cross-modal methods. Our Point-PQAE, on the other hand, only consumes unlabeled point cloud data, making it more easily extendable. Additionally, incorporating image data could increase computational requirements.

3) Joint-MAE and PiMAE cannot be called cross-reconstruction methods, unlike our Point-PQAE, because:

- Recall that cross-reconstruction methods require two components: decoupled views and a cross-reconstruction framework. In cross-reconstruction, decoupled views are obtained through independent augmentations, achieving significant diversity between views, and the cross-reconstruction framework relies on information from view 1 to mandatorily reconstruct view 2.
- The paired 3D and 2D views used by Joint-MAE and PiMAE cannot be considered isolated or decoupled views. Take PiMAE as an example: the image is merely a render from a specific camera pose of the point cloud. No augmentations can be applied to either of these views (as discussed in Section 4 of the PiMAE paper), so diversity between views cannot be achieved.
- Cross-view knowledge is used as auxiliary, not mandatory, in these two methods. If either the 3D or 2D data is removed, reconstruction can still be achieved, which turns into the case in MAE [9] or Point-MAE [14]. However, a cross-reconstruction framework should mandatorily rely on view 1 to reconstruct view 2, as in our Point-PQAE, SiamMAE [7], and CropMAE [4]. For instance, in Joint-MAE, 3D information is used as auxiliary for 2D MAE (or vice versa), and a cross-reconstruction loss (specifically, cross-modal reconstruction loss) is added to the 2D-3D output.

Thus, it is more appropriate to refer to these methods as cross-modality self-reconstruction methods.

3. Discussion on the view-relative positional embedding and positional query

Relative Positional Embedding (RPE) methods. To better position the View-Relative Positional Embedding (VRPE) proposed by us for point cloud cross-view reconstruction, we discuss the difference between our VRPE and existing RPE methods. In the fields of Natural Language Processing (NLP) and 2D vision, RPE techniques have been widely adopted [17, 20, 21]. For instance, Rotary Positional Embedding (RoPE) [18] is an emerging RPE technique gaining traction in the realm of large language models (LLMs). RoPE integrates rotational transformations to encode relative token positions, enabling more efficient extrapolation over unseen sequences. iRPE [20] first reviews existing relative position encoding methods, and then proposes new RPE methods dedicated to 2D images. The work [16] investigates the potential problems in Shaw-RPE and XL-RPE, which are the most representative and prevalent RPEs, and proposes two novel RPEs called LRHC-RPE and GCDF-RPE. Generally, in NLP and 2D vision, RPE captures the relative distances or orientations between tokens or pixels to enhance the model's capacity to understand relationships between paired elements. This approach often leads to improved generalization, especially when handling out-of-distribution data.

In contrast, our proposed VRPE is designed with a view- or instance-based focus, rather than a token-based one. Rather than capturing relationships between individual tokens or pixels, our VRPE encodes the positional relationships between two decoupled views. Our approach is not aimed at improving extrapolation or generalization. Instead, it is tailored to model the geometric and contextual information between different views to facilitate accurate cross-view reconstruction. This shift in focus makes our VRPE fundamentally distinct from existing RPE methods, highlighting the importance of carefully distinguishing our approach from existing RPE techniques.

Related Positional Query (PQ) methods. The positional query is also used in 2D self-supervised learning (SSL) and AI-generated content (AIGC). PQCL [25] pioneered the introduction of positional query, aiming to represent geometric relationships between multiple cropped views. PQDiff [26] advanced this concept by devising a contiguous relative positional query module, applying it to image outpainting to achieve arbitrary location and contiguous expansion factor outpainting. Positional query has also found applications in 2D segmentation tasks. For example, DFPQ [8] generates positional queries dynamically by

leveraging cross-attention scores from the previous decoder block and the positional encodings of the image features, which together enhance the effectiveness of semantic segmentation. Our method, however, distinguishes itself from these existing positional query approaches by focusing on the 3D world, which presents significantly greater complexity (one more dimension) and challenges compared to 2D image domains. By leveraging the obtained VRPE to query the target view from the source view, our PQ technique successfully achieves decoupled view reconstruction.

4. Additional experimental details

Training details. We utilize ShapeNet [1] as our pre-training dataset, which comprises a curated collection of 3D CAD object models, featuring 51K unique models across 55 common categories. The pre-training process spans 300 epochs, employing a cosine learning rate schedule [12] starting at 5e-4, with a warm-up period of 10 epochs. We use the AdamW optimizer [13] and a batch size of 128. All experiments are conducted on a single GPU *i.e.*, RTX 3090 (24GB). For further training details including pre-training and finetuning, refer to Tab. 3. During the pre-training of our Point-PQAE on ShapeNet, we apply rotation to the input point cloud following ReCon [15], followed by generating decoupled views from the augmented point cloud.

Finetuning evaluation protocol. For classification tasks on ScanObjectNN and ModelNet40, as well as few-shot learning on ModelNet40, we adopt three evaluation protocols, following [3, 15], to assess both the transferability of learned representations (FULL) and the quality of frozen features (MLP-LINEAR, MLP-3). The protocols are as follows:

(a) FULL: Fine-tuning the pre-trained model by updating both the backbone and the classification head.
(b) MLP-LINEAR: Fine-tuning by updating only the classification head, which consists of a single-layer linear MLP.
(c) MLP-3: Fine-tuning by updating only the parameters of a three-layer non-linear MLP classification head (which is structured the same as in FULL).

5. Additional ablation study

Integrate Positional Query (PQ) scheme into knowledge distillation. The knowledge distillation [10] typically involves inputting the same instance into both the student model and the frozen teacher model, then maximizing the mutual agreement between their outputs to distill knowledge from the teacher to the student. Our positional query block can be seamlessly integrated into knowledge distillation, allowing for cross-view distillation rather than being confined to distillation within the same view. For example, view 1 is fed to the student, view 2 is fed to the teacher, and a positional query block is added after the backbone to model relative relations and recover the latent representation of view 2. We conduct experiments on distilling the pre-trained model ReCon [15], and the results are reported in Tab. 4, indicating that our PQ scheme successfully learns knowledge from the ReCon teacher and performs much better than the baseline. It shows that the PQ scheme can be easily utilized as a plug-in tool for knowledge distillation.

Reconstruction loss function. Tab. 5 shows the performance of Point-PQAE using different reconstruction loss functions: cosine similarity loss (cos), $l1$-form Chamfer distance [5] (CD-$l1$), and the $l2$-form Chamfer distance (CD-$l2$). The results show the CD-$l2$ is more suitable for Point-PQAE.

Siamese loss function. The generative pre-training task designed by us is naturally a siamese structure and we get the form of $\mathcal{L}_{cross} = \mathcal{L}_{2\to1} + \mathcal{L}_{1\to2}$ as stated in **??**. We analyze the benefit of the siamese loss function by doing an ablation study with loss functions $\mathcal{L}_{cross} = \mathcal{L}_{2\to1} + \mathcal{L}_{1\to2}$ or $\mathcal{L}_{2\to1}$ only. The Tab. 6 presents the experiment results. It shows this siamese loss function contributes to the performance of our Point-PQAE and brings accuracy gain.

Minimum crop ratio. The minimum crop ratio r_m is important for the proposed point cloud crop mechanism. We conduct experiments to analyze the effect of minimum random crop ratios on the performance. The results are reported in Fig. 1. The results show that 0.6 is the best crop ratio for our Point-PQAE. When the ratio is too low, the model struggles to extract sufficient relevant information from the cropped view for effective cross-reconstruction. Conversely, excessively high ratios make the task too straightforward, hindering the model from learning robust representations.

Definition of views and parts in our work. We emphasize the importance of distinguishing between parts and views to understand the significance of decoupled view generation and our cross-view reconstruction method. We define the following:

- Without independently applying augmentations after cropping, the relative relationships between the cropped **parts** remain fixed.
- However, by performing view decoupling, the relative relationships between parts become more diverse, and we define these as **views**.

Existing self-reconstruction methods generally focus on cross-part reconstruction (e.g., block masking in Point-MAE [14]). In contrast, cross-view reconstruction (ours) significantly outperforms cross-part reconstruction, as demonstrated in the main paper Table 4, where line 4 outperforms line 1.

Table 3. Training details for pretraining and downstream fine-tuning.

Config	**ShapeNet**	**ScanObjectNN**	**ModelNet**	**ShapeNetPart**	**S3DIS**
optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
learning rate	5e-4	2e-5	1e-5	2e-4	2e-4
weight decay	5e-2	5e-2	5e-2	5e-2	5e-2
learning rate scheduler	cosine	cosine	cosine	cosine	cosine
training epochs	300	300	300	300	60
warmup epochs	10	10	10	10	10
batch size	128	32	32	16	32
drop path rate	0.1	0.2	0.2	0.1	0.1
number of points	1024	2048	1024	2048	2048
number of point patches	64	128	64	128	128
point patch size	32	32	32	32	32
augmentation	Rotation	Rotation	Scale&Trans	-	-
GPU device	RTX 3090	RTX 3090	RTX 3090	RTX 3090	RTX 3090

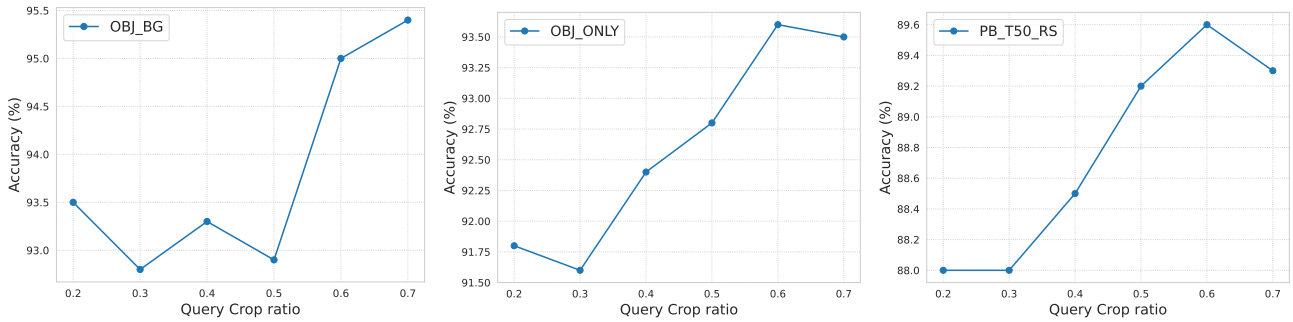

Figure 1. Ablation study on different minimum crop ratios r_m, where the results (%) of three variants: OBJ_BG, OBJ_ONLY, PT_T50_RS on ScanObjectNN are reported.

Table 4. Integrate PQ into distillation. Results on ScanobjectNN (%) are reported.

Type	OBJ_BG	OBJ_ONLY	PB_T50_RS
Train from scratch	83.0	84.0	79.1
PQ distillation	93.5	91.9	88.5

Table 5. Reconstruction loss function. The default setting is marked in gray.

Loss Function	OBJ_BG	OBJ_ONLY	PB_T50_RS
cos	90.5	89.8	85.2
CD-$l1$	93.1	91.7	89.4
CD-$l2$	95.0	93.6	89.6

Table 6. Siamese loss. The default setting is marked in gray.

Loss Function	OBJ_BG	OBJ_ONLY	PB_T50_RS
$\mathcal{L}_{2\to1}$	93.4	92.4	89.2
$\mathcal{L}_{2\to1} + \mathcal{L}_{1\to2}$	95.0	93.6	89.6

6. Limitations and future work

Point-PQAE is a novel cross-reconstruction generative learning paradigm that differs significantly from previous self-reconstruction methods, enabling more diverse and challenging pre-training. Point-MAE [14] pioneered the self-reconstruction paradigm in the point cloud self-supervised (SSL) learning field and variant optimizations are well explored, *e.g.*, cross-modal [3, 6, 15], masking strategy [24], and hierarchical architecture [23, 24]. **Compared to the well-studied self-reconstruction, cross-reconstruction remains significantly under-explored.** As the initial venture into cross-reconstruction, our Point-PQAE opens a new avenue for advancement in point cloud SSL. However, the model employs a vanilla transformer architecture and is constrained to single-modality knowledge. This architecture may not be optimally suited for cross-reconstruction tasks. Furthermore, the limited size of the available 3D point cloud datasets—due to the challenges in data collection—restricts the broader applicability of our single-modality approach. Future work could explore the

integration of knowledge from additional modalities or the development of more efficient and appropriate architectures for the cross-reconstruction paradigm.

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. 4

[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. 3

[3] 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. 4, 5

[4] Alexandre Eymaël, Renaud Vandeghen, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Efficient image pre-training with siamese cropped masked autoencoders. In *European Conference on Computer Vision*, pages 348–366. Springer, 2025. 2, 3

[5] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 4

[6] Ziyu Guo, Renrui Zhang, Longtian Qiu, 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, 3, 5

[7] Agrim Gupta, Jiajun Wu, Jia Deng, and Fei-Fei Li. Siamese masked autoencoders. *Advances in Neural Information Processing Systems*, 36:40676–40693, 2023. 2, 3

[8] Haoyu He, Jianfei Cai, Zizheng Pan, Jing Liu, Jing Zhang, Dacheng Tao, and Bohan Zhuang. Dynamic focus-aware positional queries for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11299–11308, 2023. 3

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 4

[11] Xin Li, Peng Li, Zeyong Wei, Zhe Zhu, Mingqiang Wei, Junhui Hou, Liangliang Nan, Jing Qin, Haoran Xie, and Fu Lee Wang. Cross-bert for point cloud pretraining. *arXiv preprint arXiv:2312.04891*, 2023. 1, 2

[12] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 4

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[14] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European conference on computer vision*, pages 604–621. Springer, 2022. 1, 2, 3, 4, 5

[15] 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. 4, 5

[16] Anlin Qu, Jianwei Niu, and Shasha Mo. Explore better relative position embeddings from encoding perspective for transformer models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2989–2997, 2021. 3

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 3

[18] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 3

[19] 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

[20] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10033–10041, 2021. 3

[21] Zhilin Yang. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019. 3

[22] Yaohua Zha, Huizhen Ji, Jinmin Li, Rongsheng Li, Tao Dai, Bin Chen, Zhi Wang, and Shu-Tao Xia. Towards compact 3d representations via point feature enhancement masked autoencoders. *arXiv preprint arXiv:2312.10726*, 2023. 1, 2

[23] 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. *Advances in neural information processing systems*, 35:27061–27074, 2022. 5

[24] 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, 5

[25] Shaofeng Zhang, Qiang Zhou, Zhibin Wang, Fan Wang, and Junchi Yan. Patch-level contrastive learning via positional query for visual pre-training. In *ICML*, 2023. 3

[26] Shaofeng Zhang, Jinfa Huang, Qiang Zhou, Fan Wang, Jiebo Luo, Junchi Yan, et al. Continuous-multiple image outpaint-

ing in one-step via positional query and a diffusion-based approach. In *ICLR*, 2024. 3

[27] Xiangdong Zhang, Shaofeng Zhang, and Junchi Yan. Pcp-mae: Learning to predict centers for point masked autoencoders. *Advances in Neural Information Processing Systems*, 37:80303–80327, 2024. 1, 2