DAP-MAE: Domain-Adaptive Point Cloud Masked Autoencoder for Effective Cross-Domain Learning

Ziqi Gao[1,3,4,*], Qiufu Li[2,3,4,*], Linlin Shen[2,3,4,†]
[1]School of Computer Science & Software Engineering, Shenzhen University
[2]School of Artificial Intelligence, Shenzhen University
[3]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University
[4]Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen University
gaoziqi2023@email.szu.edu.cn, {liqiufu, llshen}@szu.edu.cn

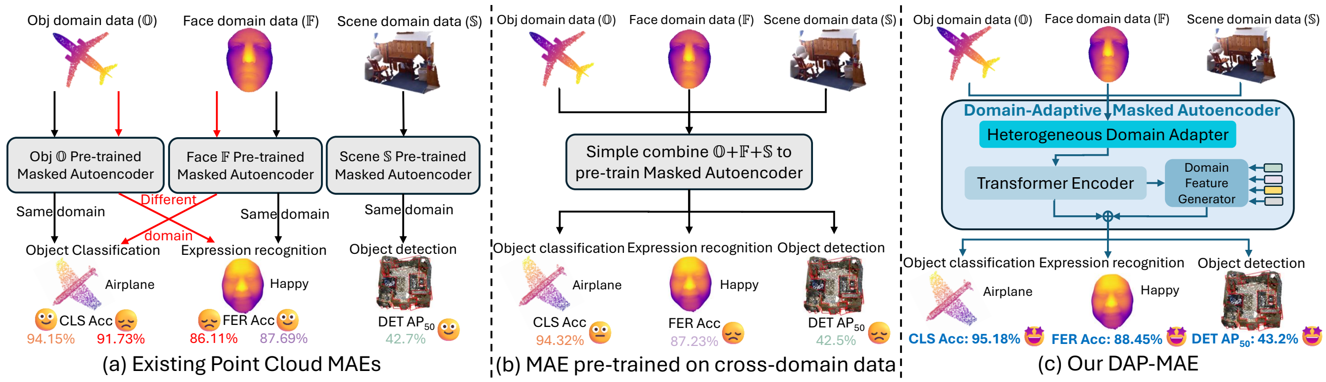

Figure 1. (a) **Existing point cloud MAEs** are typically pre-trained and fine-tuned within a single domain, leading to performance degradation when applied to other tasks in different domains. (b) Pre-training MAE directly on cross-domain data could also lead to performance decrease due to the misinterpretation of out-of-domain information. (c) **Our DAP-MAE** collaboratively learns from point cloud data across various domains using the heterogeneous domain adapter and adopts the domain feature generator to extract diverse domain features, allowing the model to achieve high performance across multiple tasks with one pre-training in a single modality.

Abstract

Compared to 2D data, the scale of point cloud data in different domains available for training, is quite limited. Researchers have been trying to combine these data of different domains for masked autoencoder (MAE) pre-training to leverage such a data scarcity issue. However, the prior knowledge learned from mixed domains may not align well with the downstream 3D point cloud analysis tasks, leading to degraded performance. To address such an issue, we propose the ***Domain-Adaptive Point Cloud Masked Autoencoder (DAP-MAE)****, an MAE pre-training method, to adaptively integrate the knowledge of cross-domain datasets for general point cloud analysis. In DAP-MAE, we design a heterogeneous domain adapter that utilizes an adaptation mode during pre-training, enabling the model to comprehensively learn information from point clouds across different domains, while employing a fusion mode in the fine-tuning to enhance point cloud features. Meanwhile, DAP-MAE incorporates a domain feature generator to guide the adaptation of point cloud features to various downstream tasks. With only one pre-training, DAP-MAE achieves excellent performance across four different point cloud analysis tasks, reaching 95.18% in object classification on ScanObjectNN and 88.45% in facial expression recognition on Bosphorus. The code will be released at https://github.com/CVI-SZU/DAP-MAE*

1. Introduction

Point cloud analysis is gaining increasing attention in the fields of autonomous driving, robotics, and augmented/virtual reality, etc. It involves tasks such as point cloud-based object classification, part segmentation, facial expression recognition, and object detection, all of them require training feature extraction models that can effectively represent the geometric information of 3D point clouds.

*Equal contribution.
†Corresponding author.

However, compared to 2D data, the collection and annotation of 3D point clouds still require substantial resources, resulting in small-scale labeled point cloud datasets, which limits the performance of supervised learning methods [26, 32, 33]. Self-supervised learning [25, 30, 49] offers a promising alternative for the model training, utilizing unlabeled 3D point clouds and pre-training strategies to learn the features.

The self-supervised learning method based on masked autoencoders (MAE) can fully leverage the geometric information of 3D point clouds, enhancing the performance of downstream tasks, and various MAE-based methods for 3D point cloud analysis have been developed, such as Point-MAE [30], PiMAE [3], and 3DFaceMAE [8], etc. As Fig. 1 shows, when these methods are applied to downstream tasks including object classification, facial expression recognition, and indoor object detection, they require separate pre-training on the data from the same domain as the tasks, which leads to redundant consumption of pre-training and does not fully utilize the existing point cloud data from different domains. However, if the MAE is simply pre-trained using 3D point cloud data from multiple different domains, the downstream tasks may interpret the out-of-domain information learned by the feature model as interference noises, thereby reducing the task performance.

In this paper, to comprehensively utilize the 3D point cloud data from different domains, we designed a domain-adaptive point cloud masked autoencoder (DAP-MAE). DAP-MAE employs a heterogeneous domain adapter (HDA) and a domain feature generator (DFG) to collaboratively learn point cloud data from various domains, enabling the model to adaptively perform different downstream tasks with just one pre-training in a single modality, as Fig. 1(c) shows. Specifically, during the pre-training, the HDA employs an adaptation mode to process masked point clouds from different domains using three parallel MLPs, enabling the model to collaboratively learn cross-domain information. During the fine-tuning, the HDA utilizes a fusion mode to integrate the point cloud data output from the three MLPs and enhance the point cloud features extracted by the MAE. Meanwhile, the DFG is pre-trained to extract relevant domain features of the point cloud using a contrastive loss to guide the adaptation of the point cloud features to various downstream tasks in the fine-tuning. Our main contributions are summarized as follows:

- For the first time, we propose a domain-adaptive point cloud masked autoencoder, DAP-MAE, that can collaboratively utilize point clouds from different domains for various downstream tasks with just one pre-training in a single modality.
- In DAP-MAE, we design a heterogeneous domain adapter to address collaborative learning on cross-domain data and a domain feature generator extracting various domain features to guide the downstream tasks.
- The experiments demonstrate that DAP-MAE achieves SOTA results in various point cloud analysis tasks including object classification, facial expression recognition, part segmentation, and object detection, surpassing or comparable to cross-modal methods.

2. Related Work

2.1. 3D point cloud analysis with deep learning

Deep learning-based point cloud analysis has been applied to various downstream tasks. Early point cloud analysis methods focused on object analysis tasks like object classification and part segmentation [11, 22, 26, 32, 33, 36, 46]. PointNet [32] pioneered a direct approach to consuming unordered point sets using symmetric functions to capture global features, and its extension, PointNet++ [33] introduced a hierarchical set abstraction that captures local information. Built on deeper research in object analysis tasks, some methods have shifted their focus to 3D face analysis tasks [9, 17, 20, 52]. Among these, PointFace [17] introduced the first 3D face recognition method based on PointNet++ and also proposed a transformer-based variant called PointFaceFormer [9]. Meanwhile, DrFER [20] put forward the first 3D facial expression recognition method. At the same time, point cloud analysis has also advanced toward 3D scene understanding [27, 34, 39, 63]. 3DETR [27] employs a fully transformer-based design to capture both local geometry and global context for enhancement of 3D object detection accuracy. Despite their strong performance in individual tasks, these methods differ in structure, a unified model is thus required for all the downstream tasks. Meanwhile, by integrating the power of Transformers with self-supervised learning on unlabeled point clouds, Point-MAE [30]—the first MAE-based method under a standard transformer framework—has emerged as a promising way to address this limitation.

2.2. MAE-based point cloud analysis

With Point-MAE [30] pioneering the masked reconstruction paradigm in 3D vision, numerous generative approaches with cross-modal distillation have since been proposed. ACT [7] transfers a Transformer pre-trained on other modalities to 3D point clouds and distills knowledges of other modalities into the point cloud Transformer. ReCon [35] unified generative modeling and contrastive learning through ensemble distillation guided by cross-modal teachers. Since cross-modal methods require pre-training on additional modalities, such a framework require high training costs, and highly rely on the performance of the teacher model. As a result, many approaches in 2024 have shifted back to improving single-modal strategies [23, 57]. Also, some methods such as 3DFaceMAE [8] and Li *et*

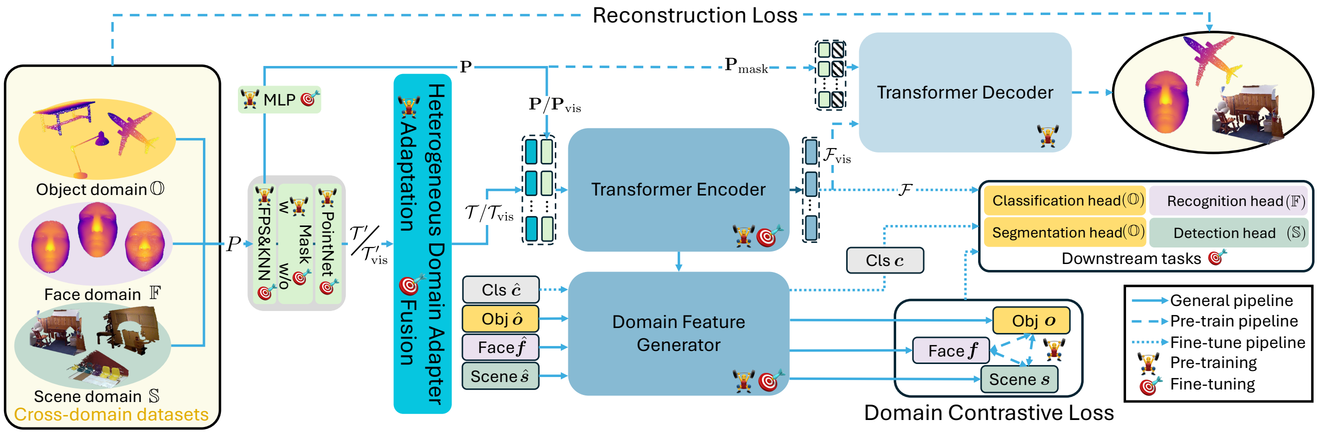

Figure 2. The pipeline of DAP-MAE. During the pre-training, the heterogeneous domain adapter (HDA) is set to adaptation mode to enhance the training of the Transformer-based MAE on cross-domain point cloud datasets, while the domain feature generator (DFG) is trained to extract domain features from the point clouds. In the fine-tuning phase, HDA is set to fusion mode, making comprehensive use of the feature extraction capabilities of the model trained on cross-domain data; meanwhile, the domain features extracted by DFG are used to guide the adaptation of point cloud features for downstream tasks.

al. [21] have applied MAE-based methods to face analysis tasks. In the field of scene understanding, PiMAE [3] and GDMAE [51] have adopted MAE-based methods for indoor and outdoor scene understanding tasks, respectively. However, existing methods have not thoroughly explored the relationships between different domains in current 3D point cloud datasets, resulting in many approaches performing well only on downstream tasks closely aligned with their pre-training datasets.

3. Preliminary

The point cloud analysis addressed in this paper involves point clouds from three different data domains.

- Object domain, $\mathbb{O}$. The tasks of 3D object classification and part segmentation utilize point cloud datasets from the object domain, such as ShapeNet [2], ModelNet [48], and ScanObjectNN [42].
- Face domain, $\mathbb{F}$. The tasks of 3D face recognition and facial expression recognition use point cloud datasets from the face domain, such as FRGCv2 [31], Bosphorus [38], and BU3DFE [54].
- Scene domain, $\mathbb{S}$. The tasks of 3D object detection use point cloud datasets from the scene domain, such as ScanNet [6] and S3DIS [1].

The existing self-supervised point cloud analysis methods struggle to effectively utilize point cloud data from different domains. If the point clouds from different domains are used to pre-train feature extraction models, the performance of downstream tasks will significantly decline. As Fig. 1(a) shows, when we pre-train ReCon-SMC [35] on object data ($\mathbb{O}$) and then transfer it to object classification ($\mathbb{O}$), or pre-train it on face data ($\mathbb{F}$) and then transfer it to facial expression recognition ($\mathbb{F}$), their performance are 94.15% and 87.69%, respectively. However, when pre-training and the downstream task are conducted across different domains, such as pre-training on face data ($\mathbb{F}$) and transferring to object classification ($\mathbb{O}$), or pre-training on object data ($\mathbb{O}$) and transferring to facial expression recognition ($\mathbb{F}$), the performance drops to 91.73% and 86.11%, respectively. Furthermore, directly combine point clouds from multiple domains together for the pre-training also struggles to improve the performance of downstream tasks. As Fig. 1(b) shows, when additional face data ($\mathbb{F}$) and scene data ($\mathbb{S}$) are sequentially added during the pre-training, the object classification ($\mathbb{O}$) performance only slightly improved by 0.17%, resulting in an accuracy of 94.32%. Moreover, the performance on facial expression recognition and object detection tasks even degrade to 87.23% and 42.5% respectively.

To fully exploit the information from data across different domains, this paper designs a domain-adaptive point cloud Masked Autoencoder, which requires only a single pre-training on point cloud data from multiple domains to achieve improved point cloud analysis accuracy on various downstream tasks.

4. Method

As Fig. 2 shows, in this section, we first introduce the overall pipeline of our DAP-MAE, followed by detail description of the heterogeneous domain adapter (HDA) and domain feature generator (DFG).

4.1. The pipeline of DAP-MAE

Datasets. To address the multiple point cloud analysis tasks including object classification, facial expression recognition, part segmentation, and object detection, the proposed DAP-MAE is pre-trained on point clouds from

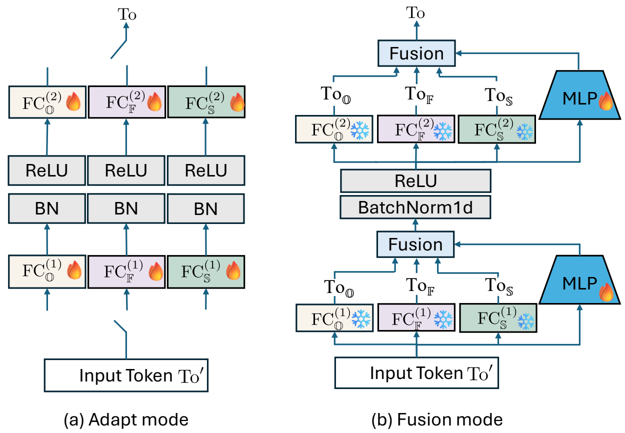

Figure 3. Illustration of the heterogeneous domain adapter (HDA), which consists of three parallel MLPs and two modes. (a) Adaptation mode. In the pre-training, HDA adaptively selects a MLP to process the input token based on its domain. (b) Fusion mode. In the fine-tuning, HDA freezes the parameters in the three MLPs and trains two additional MLPs to generate fusion coefficients, which are used to linearly fuse the output by the three MLPs.

three different domains, i.e., object $\mathbb{O}$, face $\mathbb{F}$, and scene $\mathbb{S}$, while, in the fine-tuning phase, it is refined by the point clouds from the corresponding domain based on the specific downstream task.

Tokenization. For a point cloud P from domain $\mathbf{d}$, we first use farthest point sampling (FPS) and k-nearest neighbors (KNN) to partition it into a set of patches $\mathcal{P}$, where each patch in $\mathcal{P}$ contains a center point and k neighboring points. In the pre-training, with a ratio of m, we randomly divide the patch set $\mathcal{P}$ into visible and masked parts,

$$\mathcal{P}_{\text{vis}}, \mathcal{P}_{\text{mask}} = \text{Mask} \circ \text{KNN} \circ \text{FPS}(P). \tag{1}$$

Before feeding into the Transformer-based MAE, we tokenize the patches using PointNet [32] and the proposed heterogeneous domain adapter (HDA),

$$\mathcal{T} = [\mathcal{T}_{\text{vis}}, \mathcal{T}_{\text{mask}}] = \text{HDA} \circ \text{PointNet}(\mathcal{P}_{\text{vis}}, \mathcal{P}_{\text{mask}}), \tag{2}$$

and generate the position embedding using the center point of each patch by an MLP,

$$\mathbf{P} = [\mathbf{P}_{\text{vis}}, \mathbf{P}_{\text{mask}}] = \text{MLP}(\mathcal{P}_{\text{vis}}, \mathcal{P}_{\text{mask}}). \tag{3}$$

Reconstruction. In the pre-traing, we train the Transformer-based MAE on the visible data and reconstruct point clouds using a self-supervised manner. The Transformer encoder extracts the feature of the point cloud from the visible $\mathcal{T}_{\text{vis}}$ and $\mathbf{P}_{\text{vis}}$,

$$\mathcal{F}_{\text{vis}} = \text{Encoder}(\mathcal{T}_{\text{vis}}, \mathbf{P}_{\text{vis}}), \tag{4}$$

and the decoder reconstructs the masked patches,

$$\hat{\mathcal{P}}_{\text{mask}} = \text{Decoder}(\mathcal{F}_{\text{vis}}, \mathbf{P}_{\text{vis}}, \mathbf{P}_{\text{mask}})). \tag{5}$$

The reconstruction in the MAE is supervised by a chamfer distance loss mutually comparing the patches in $\mathcal{P}_{\text{mask}}$ and $\hat{\mathcal{P}}_{\text{mask}}$,

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathcal{P}_{\text{mask}}|} \sum_{\text{Pa} \in \mathcal{P}_{\text{mask}}} \min_{\text{Pa}' \in \hat{\mathcal{P}}_{\text{mask}}} \|\text{Pa} - \text{Pa}'\|_2^2 + \frac{1}{|\hat{\mathcal{P}}_{\text{mask}}|} \sum_{\text{Pa} \in \hat{\mathcal{P}}_{\text{mask}}} \min_{\text{Pa}' \in \mathcal{P}_{\text{mask}}} \|\text{Pa} - \text{Pa}'\|_2^2. \tag{6}$$

Domain contrasting. To extract domain features that guide the downstream tasks, we train the domain feature generator (DFG) in the pre-training to decompose the feature $\mathcal{F}_{\text{vis}}$ into the domain feature $\boldsymbol{d}$,

$$\boldsymbol{d} = \text{DFG}(\mathcal{F}_{\text{vis}}), \tag{7}$$

In each iteration of the pre-training, DFG extracts the domain features $[\boldsymbol{d}_i]_{i=1}^B$ of a batch of cross-domain point clouds $[P_i]_{i=1}^B$ from domains $[\mathbf{d}_i]_{i=1}^B$, $\mathbf{d}_i \in \{\mathbb{O}, \mathbb{F}, \mathbb{S}\}$, and it is trained using a contrastive loss,

$$\mathcal{L}_{\text{con}} = \sum_{i=1}^{B} \sum_{j=1, j\neq i}^{B} l(\boldsymbol{d}_i, \boldsymbol{d}_j), \tag{8}$$

$$l(\boldsymbol{d}_i, \boldsymbol{d}_j) = \begin{cases} 1 - \cos(\boldsymbol{d}_i, \boldsymbol{d}_j), & \mathbf{d}_i = \mathbf{d}_j \\ \max(0, \cos(\boldsymbol{d}_i, \boldsymbol{d}_j) - a), & \mathbf{d}_i \neq \mathbf{d}_j, \end{cases} \tag{9}$$

and a is a margin constant.

Fine-tuning. During fine-tuning, for a point cloud P from a specialized domain, the Transformer encoder extracts its features from the tokens $\mathcal{T}$, and the DFG utilizes these features to decompose domain features while additionally extracting class features $\boldsymbol{c}$ with class token,

$$\mathcal{F} = \text{Encoder}(\mathcal{T}, \mathbf{P}), \tag{10}$$

$$\boldsymbol{c}, \boldsymbol{d} = \text{DFG}(\mathcal{F}), \tag{11}$$

which are then fed into a downstream task head together, and the head is trained using its customized loss.

4.2. Heterogeneous Domain Adapter

The Heterogeneous Domain Adapter (HDA) consists of three parallel MLPs, which, in the pre-training and fine-tuning phases, process point cloud data using different modes. As Fig. 3 shows, in the pre-training, HDA employs an adaptation mode, processing point cloud data from the three different domains using the different MLPs respectively, which synergistically leverages cross-domain data to enhance the feature extraction of encoder in MAE. In the fine-tuning, HDA adopts a fusion mode, where the three MLPs simultaneously process the point cloud data and extract its features from their fused results.

Adaptation mode. Suppose a point could P from domain $\mathbf{d}, \mathbf{d} \in \{\mathbb{O}, \mathbb{F}, \mathbb{S}\}$. In the pre-training, for its visible patch set $\mathcal{P}_{\text{vis}}$, PointNet first produces its initial tokens,

$$\mathcal{T}'_{\text{vis}} = \text{PointNet}(\mathcal{P}_{\text{vis}}) = \{\text{PointNet}(\text{Pa}) : \text{Pa} \in \mathcal{P}_{\text{vis}}\}. \tag{12}$$

Then, HDA generates its tokens using one of three parallel MLPs, $\{\text{MLP}_{\mathbf{d}} : \mathbf{d} \in \{\mathbb{O}, \mathbb{F}, \mathbb{S}\}\}$, according to the domain of $P \in \mathbf{d}$,

$$\begin{aligned} \mathcal{T}_{\text{vis}} &= \text{HDA}(\mathcal{T}'_{\text{vis}}) = \text{MLP}_{\mathbf{d}}(\mathcal{T}'_{\text{vis}}) \\ &= \{\text{To} = \text{MLP}_{\mathbf{d}}(\text{To}') : \text{To}' \in \mathcal{T}'_{\text{vis}}\}. \end{aligned} \tag{13}$$

In HDA, each of three MLPs contains two linear full connection (FC) layers, $\text{FC}_{\mathbf{d}}^{(1)}$ and $\text{FC}_{\mathbf{d}}^{(2)}$, with a BN and ReLU, and for every initial token To' in $\mathcal{T}'_{\text{vis}}$,

$$\text{To} = \text{FC}_{\mathbf{d}}^{(2)} \circ \text{ReLU} \circ \text{BN} \circ \text{FC}_{\mathbf{d}}^{(1)}(\text{To}') \in \mathcal{T}_{\text{vis}}. \tag{14}$$

Fusion mode. During the fine-tuning, for the different downstream task, we respectively fine-tune the feature extraction parameters of DAP-MAE and the downstream task heads on point cloud data from the same domain. For the point cloud P from domain $\mathbf{d} \in \{\mathbb{O}, \mathbb{F}, \mathbb{S}\}$, we generate all its initial tokens using FPS, KNN, and PointNet,

$$\mathcal{T}' = \text{PointNet} \circ \text{KNN} \circ \text{FPS}(P). \tag{15}$$

In the fine-tuning, as Fig. 3 shows, HDA freezes its parameters and processes each initial token with the three parallel MLPs separately, and twice linearly fuses the MLP results. In the first fusion, for each initial token $\text{To}' \in \mathcal{T}'$ of point cloud $P \in \mathbf{d}$, HDA processes it using the first three FCs in the three parallel MLPs,

$$\text{To}_{\mathbf{d}'}^{(1)} = \text{FC}_{\mathbf{d}'}^{(1)}(\text{To}'), \ \ \forall \mathbf{d}' \in \{\mathbb{O}, \mathbb{F}, \mathbb{S}\}, \forall \text{To}' \in \mathcal{T}'. \tag{16}$$

In the fusion, the token output by the FC corresponding to the domain $\mathbf{d}$ of downstream task is taken as the primary token, while the other tokens from FCs for other domains serve as auxiliary tokens,

$$\text{To}^{(1)} = \text{To}_{\mathbf{d}}^{(1)} + \sum_{\substack{\mathbf{d}' \in \{\mathbb{O}, \mathbb{F}, \mathbb{S}\} \\ \mathbf{d}' \neq \mathbf{d}}} \alpha_{\mathbf{d}'} \text{To}_{\mathbf{d}'}^{(1)}, \tag{17}$$

where the two coefficients are calculated using one MLP on the initial token,

$$\{\alpha_{\mathbf{d}'} : \mathbf{d}' \in \{\mathbb{O}, \mathbb{F}, \mathbb{S}\}, \mathbf{d}' \neq \mathbf{d}\} = \text{MLP}^{(1)}(\text{To}'). \tag{18}$$

Then, in the second fusion, the token $\text{To}^{(1)}$ is first processed by ReLU and BN, and fused by the second three FCs of the three parallel MLPs with the similar way,

$$\text{To}_{\mathbf{d}'}^{(2)} = \text{FC}_{\mathbf{d}'}^{(2)}(\text{ReLU}(\text{BN}(\text{To}^{(1)}))), \ \mathbf{d}' \in \{\mathbb{O}, \mathbb{F}, \mathbb{S}\}, \tag{19}$$

$$\text{To} = \text{To}_{\mathbf{d}}^{(2)} + \sum_{\substack{\mathbf{d}' \in \{\mathbb{O}, \mathbb{F}, \mathbb{S}\} \\ \mathbf{d}' \neq \mathbf{d}}} \alpha_{\mathbf{d}'} \text{To}_{\mathbf{d}'}^{(2)}, \tag{20}$$

where the two coefficients are calculated using another MLP on the token $\text{To}^{(1)}$,

$$\{\alpha_{\mathbf{d}'} : \mathbf{d}' \in \{\mathbb{O}, \mathbb{F}, \mathbb{S}\}, \mathbf{d}' \neq \mathbf{d}\} = \text{MLP}^{(2)}(\text{To}^{(1)}). \tag{21}$$

4.3. Domain Feature Generator

The domain feature generator (DFG) extracts domain features from the point cloud features $\mathcal{F}_{\text{vis}}$ or $\mathcal{F}$ output by the Transformer encoder, which are used to guide the training of downstream task heads during the fine-tuning phase.

In the DFG, a class token $\hat{\boldsymbol{c}}$ and three domain tokens $\{\hat{\boldsymbol{o}}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{s}}\}$ are set. During the pre-training, based on the domain $\mathbf{d}$ of the original input point cloud P, one domain token $\hat{\boldsymbol{d}} \in \{\hat{\boldsymbol{o}}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{s}}\}$ is selected and concatenated with the class token $\hat{\boldsymbol{c}}$. Then, domain features are extracted from $\tilde{\mathcal{F}}$, $\tilde{\mathcal{F}} = \mathcal{F}_{\text{vis}}$ or $\mathcal{F}$, using a cross-attention mechanism,

$$Q = \text{FC}_q([\hat{\boldsymbol{c}}, \hat{\boldsymbol{d}}]), K = \text{FC}_k(\tilde{\mathcal{F}}), V = \text{FC}_v(\tilde{\mathcal{F}}), \tag{22}$$

$$[\boldsymbol{c}, \boldsymbol{d}] = \text{Attn}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}}\right)V. \tag{23}$$

where d is the dimension of the key vectors K and $\boldsymbol{c}, \boldsymbol{d}$ are the class feature and domain feature extracted by DFG. To be clear, the class token $\hat{\boldsymbol{c}}$ are only supervised during the fine-tuning stage to extract class feature $\boldsymbol{c}$.

5. Experiments

5.1. Pre-training

Cross-domain dataset. The pre-training of DAP-MAE was conducted on a cross-domain point cloud dataset, which involve three datasets from different domains: ShapeNet ($\mathbb{O}$), FRGCv2 ($\mathbb{F}$), and S3DIS ($\mathbb{S}$). For each point cloud in the cross-domain dataset, we uniformly sample its 4,096 points for the pre-training. More details about the cross-domain dataset will be provided in the supplementary material.

5.2. Fine-tuning

The pre-trained DAP-MAE was fine-tuned on five representative datasets covering diverse downstream tasks, including object classification ($\mathbb{O}$), few-shot learning ($\mathbb{O}$), part segmentation ($\mathbb{O}$), facial expression recognition ($\mathbb{F}$), and 3D object detection ($\mathbb{S}$). For each task, we selected widely used benchmark datasets. These include ScanObjectNN [42] for object classification, ModelNet40 [48] for few-shot learning classification, ShapeNetPart [2] for part segmentation, BU-3DFE [54] and Bosphorus [38] for facial expression recognition, and ScanNetV2 [6] for indoor scene 3D object detection. More detailed descriptions and implementation settings can be found in the supplementary material.

Object classification ($\mathbb{O}$). The left side of Tab. 1 shows the classification accuracy (%) on ScanObjectNN. Our DAP-MAE outperforms the latest single-modal self-supervised

Table 1. Object Classification results on the ScanObjectNN and Face Expression Recognition results on BU3DFE and Bosphorus. PM: pre-trained modality, *PC*: point cloud, *I*: image, *T*: text.

Object Classification					Facial Expression Recognition		
Method	PM	ScanObjectNN			Method	BU3DFE	BOS
		OBJ_BG	OBJ_ONLY	PB_T50_RS			
Supervised Learning					2D+3D		
PointNet [32] CVPR'17	*PC*	73.3	79.2	68.0	DA-CNN [64]	88.35	-
PointNet++ [33] NeruIPS'17	*PC*	84.3	84.3	77.9	Jan *et al.* [16]	88.54	-
DGCNN [46] TOG'19	*PC*	82.6	86.2	78.1	Zhu *et al.* [50]	88.75	-
PointCNN [22] NeruIPS'18	*PC*	86.1	85.5	78.5	CM-CNN [28]	88.91	85.16
SimpleView [10] ICML'21	*PC*	-	-	80.5±0.5	FE3DNet[24]	89.05	89.28
MVTN [13] ICCV'21	*PC*	92.6	-	82.8	FA-CNN [18]	89.11	-
PointMLP [26] ICLR'22	*PC*	-	-	85.4±0.3	Oyedotun *et al.* [29]	89.31	-
PCT [11] CVM'21	*PC*	-	-	83.4	Jiao *et al.* [19]	89.72	83.63
PointNeXt [36] NeruIPS'22	*PC*	-	-	87.7±0.4	FFNet-M [40]	89.82	87.65
P2P-HorNet [45] NeruIPS'22	*PC*	-	-	89.3	AFNet-M [41]	90.08	88.31
SFR [56] ICASSP'23	*PC*	-	-	87.8	Cmanet [65]	90.24	89.36
Self-Supervised Learning					3D		
Point-BERT [55] CVPR'22	*PC*	87.43	88.11	83.07	Jan *et al.* [16]	81.83	-
Point-MAE [30] ECCV'22	*PC*	90.02	88.29	85.18	CM-CNN [28]	80.11	77.82
Point-M2AE [58] NeruIPS'22	*PC*	91.22	88.81	86.43	Zhen *et al.* [61]	84.50	-
PointGPT-S [4] NeruIPS'23	*PC*	91.60	90.00	86.90	FLM-CNN [5]	86.67	-
PointDif [62] CVPR'24	*PC*	91.91	93.29	87.61	Zhu *et al.* [50]	87.19	-
Mamba3D+Point-MAE [14] ACM MM'24	*PC*	93.12	92.08	88.20	FFNet-M [40]	87.28	82.86
Point-FEMAE [57] AAAI'24	*PC*	**95.18**	93.29	90.22	Oyedotun *et al.* [29]	84.72	-
ReCon-SMC [35] ICML'23 (baseline)	*PC*	94.15	93.12	89.73	Yang *et al.* [53]	84.80	77.50
DAP-MAE(Ours)	*PC*	**95.18**	**93.45**	**90.25**	FE3DNet[24]	85.20	83.55
ACT [7] ICLR'23	*PC+I*	93.29	91.91	88.21	DA-CNN [64]	87.69	-
Joint-MAE [12] IJCAI'23	*PC+I*	90.94	88.86	86.07	Cmanet [65]	84.03	81.25
I2P-MAE [59] CVPR'23	*PC+I*	94.14	91.57	90.11	AFNet-M [41]	86.97	82.06
TAP [47] ICCV'23	*PC+I*	90.36	89.50	85.67	DrFER [20]	89.15	86.77
ReCon-full [35] ICML'23	*PC+I+T*	95.18	93.63	90.63	**DAP-MAE(Ours)**	**89.83**	**88.45**

Table 2. Results for object classification, facial expression recognition (FER), and object detection (DET). We compare Point-MAE and ReCon-SMC (both using same-domain pre-training and transfer) with their cross-domain version trained using the simple combination of different domain data, as well as our DAP-MAE. PM: pre-trained modality, *PC*: point cloud. † denotes the reproduced results.

Method	PM	object classification			FER		DET	
		ScanObjectNN			BU3DFE	BOS	ScanNetV2	
		OBJ_BG	OBJ_ONLY	PB_T50_RS			AP_{50}	AP_{25}
Point-MAE [30] ECCV'22	*PC*	90.02	88.29	85.18	88.89	86.75	-	-
ReCon-SMC [35] ICML'23 (baseline)	*PC*	94.15	93.12	89.73	89.13	87.69	42.7	63.8
$\mathbb{O}+\mathbb{F}+\mathbb{S}$ Point-MAE† [30]	*PC*	92.77	91.22	88.24	87.34	86.28	-	-
$\mathbb{O}+\mathbb{F}+\mathbb{S}$ ReCon-SMC † [35] (baseline)	*PC*	94.32	93.12	89.90	88.52	87.23	42.5	63.5
DAP-MAE (Ours)	*PC*	**95.18**	**93.45**	**90.25**	**89.83**	**88.45**	**43.2**	**64.0**

learning approaches across all three protocols and exceeds the performance of most cross-modal methods. Additionally, in Tab. 2, we compare Point-MAE and ReCon-SMC, which are both pre-trained and transferred on the same domain with their counterparts pre-trained on the cross-domain dataset. The results show that simply adding cross-domain data only gains slight improvements in object classification for ReCon-SMC, whereas our improved DAP-MAE significantly surpasses the single-modal baseline ReCon-SMC by 1.03% and its cross-domain version by 0.86%, demonstrating its effective use of cross-domain data to enhance downstream classification performance.

Facial expression recognition ($\mathbb{F}$). The rights side of Tab. 1 shows the effectiveness of our DAP-MAE on facial expression recognition. On both datasets, the experiments follow the same protocol, where 60 subjects are selected for 10-fold cross-validation. DAP-MAE achieves the highest recognition accuracy on BU-3DFE and Bosphorus compared with the latest facial expression recognition approaches. Our DAP-MAE outperforms the latest method DrFER [20] by 0.66% and 1.68% on the BU-3DFE and Bosphorus datasets, respectively. Moreover, our method also surpasses most 2D+3D multi-modal approaches and achieves performance very close to the best-performing multi-modal method. As Tab. 2 shows, when pre-trained with simple combination of $\mathbb{O}+\mathbb{F}+\mathbb{S}$, Point-MAE and ReCon-SMC still perform lower than DAP-MAE and even lower than their same-domain version. This may due to the reason that other-domain data is treated as interference noises, as previously discussed. Table 1 shows that DAP-

Table 3. Few-shot learning results on ModelNet40.

Method	5-way		10-way	
	10-shot	20-shot	10-shot	20-shot
Single-Modal Self-Supervised Learning				
Point-BERT[55] CVPR'22	94.6±3.1	96.3±2.7	91.0±5.4	92.7±5.1
Point-MAE [30] ECCV'22	96.3±2.5	97.8±1.8	92.6±4.1	95.0±3.0
Point-M2AE [58] NeruIPS'22	96.8±1.8	98.3±1.4	92.3±4.5	95.0±3.0
PointGPT- [4] NeruIPS'23	96.8±2.0	98.6±1.1	92.6±4.6	95.2±3.4
Point-FEMAE[57] AAAI'24	97.2±1.9	98.6±1.3	**94.0±3.3**	**95.8±2.8**
DAP-MAE(Ours)	**97.5±1.8**	**98.9±0.6**	93.3±3.9	95.2±2.8
Cross-Modal Self-Supervised Learning				
ACT[7] ICLR'23	96.8±2.3	98.0±1.4	93.3±4.0	95.6±2.8
Joint-MAE[12]IJCAI'23	96.7±2.2	97.9±1.8	92.6±3.7	95.1±2.6
I2P-MAE[59] CVPR'23	97.0±1.8	98.3±1.3	92.6±5.0	95.5±3.0
TAP[47] ICCV'23	97.3±1.8	97.8±1.9	93.1±2.6	95.8±1.0
ReCon-full[35] ICML'23	97.3±1.9	98.9±1.2	93.3±3.9	95.8±3.0

Table 4. Part segmentation results on the ShapeNetPart: Mean intersection over union for all classes mIoU_c (%) and all instances mIoU_I (%) for Part Segmentation. PM: pre-trained modality, *PC*: point cloud, *I*: image, *T*: text.

Method	PM	mIoU_c	mIoU_I
Supervised Learing			
PointNet[32] CVPR'17	*PC*	80.4	83.7
PointNet++[33] NeruIPS'17	*PC*	81.9	85.1
PointMLP [26] ICLR'22	*PC*	84.6	86.1
Self-Supervised Learning			
Transformer [43] NeruIPS'17	*PC*	83.4	84.7
Transformer-OcCo [44] ICCV'21	*PC*	83.4	85.1
Point-BERT [55] CVPR'22	*PC*	84.1	85.6
MaskPoint [25] ECCV'22	*PC*	84.4	86.0
Point-MAE [30] ECCV'22	*PC*	84.2	86.1
PointGPT-S[4] NeruIPS'23	*PC*	84.1	86.2
PM-MAE[23] TCSVT'24	*PC*	84.3	85.9
Mamba3D+Point-MAE [14] ACM MM'24	*PC*	83.6	85.6
DAP-MAE(Ours)	*PC*	**84.9**	**86.3**
ACT[7] ICLR'23	*PC+I*	84.7	86.1
ReCon-full[35] ICML'23	*PC+I+T*	84.9	86.4

MAE not only performs well on coarse-grained classification tasks but also excels in fine-grained classification, highlighting its capability to address the domain gap problem.

Few-shot learning ($\mathbb{C}$). Table 3 shows the results of few-shot learning, where each experiment undergoes 10 independent trials, and the average result is reported. Our DAP-MAE achieves the best performance among both single-modal and cross-modal methods, reaching 97.5% on the "5-way, 10-shot" and 98.9% on the "5-way, 20-shot" settings. For the "10-way, 10-shot" setting, our method outperforms all cross-modal methods and is only lower than Point-FEMAE. These results demonstrate that DAP-MAE, compared to other methods, maintains strong performance even with limited fine-tuning data.

Part segmentation ($\mathbb{C}$). Table 4 presents the results of object part segmentation on the ShapeNetPart [2]. Among all single-modal self-supervised methods, our DAP-MAE achieved the highest mIoU_c and mIoU_I, reaching 84.9% and 86.3%, respectively, with mIoU_c being 0.4% higher than the runner up. Additionally, DAP-MAE outperformed existing supervised methods in the part segmentation task, being slightly lower than the cross-modal method ReCon-full only in the mIoU_I. This experiment demonstrates the effectiveness of our approach in segmentation tasks.

Table 5. 3D object detection on the ScanNetV2 dataset: The detection performance using Average Precision (AP) at two different IoU thresholds of 0.50 and 0.25, i.e., AP_{50} and AP_{25} are reported. PM: pre-trained modality, *PC*: point cloud, *I*: image, *T*: text.

Method	PM	SSL	Input	AP_{50}	AP_{25}
VoteNet[34] ICCV'19	*PC*	×	xyz	33.5	58.6
PointContrast [49] ECCV'20	*PC*	✓	xyz	38.0	59.2
STRL [15] ICCV'21	*PC*	✓	xyz	38.4	59.5
RandomRooms [37] ICCV'21	*PC*	✓	xyz	36.2	61.3
DepthContrast[60] ICCV'21	*PC*	✓	xyz	-	61.3
3DETR [27] ICCV'21	*PC*	×	xyz	37.9	62.1
Point-BERT [55] CVPR'22	*PC*	✓	xyz	38.3	61.0
MaskPoint[25] ECCV'22	*PC*	✓	xyz	40.6	63.4
PiMAE[3] CVPR'23	*PC+I*	✓	xyz	39.4	62.6
ACT[7] ICLR'23	*PC+I*	✓	xyz	42.1	63.8
DAP-MAE(Ours)	*PC*	✓	xyz	**43.2**	**64.0**

3D Object Detection ($\mathbb{S}$). DAP-MAE was evaluated on ScanNetV2 [6] for the indoor 3D object detection, using the detection head of 3DETR [27]. Table 5 presents the Average Precision (AP) of different methods on the validation set, at thresholds 0.25 and 0.50. Our DAP-MAE achieved 64.0% AP_{25} and 43.2% AP_{50}, surpassing the baseline 3DETR by 5.3% and 1.9%, respectively, achieving the highest performance among all methods. Compared to PiMAE [3], a special method for 3D object detection, our DAP-MAE improves by 3.8% and 1.4% on the two metrics. Also, DAP-MAE outperforms the cross-modal approach ACT with gains of 1.1% and 0.2%, further demonstrating its effectiveness. Furthermore, as shown in Tab. 2, pre-training ReCon-SMC with cross-domain data leads to a 0.2% AP_{50} drop compared to the same-domain version, whereas DAP-MAE instead achieves a 0.5% improvement.

5.3. Ablation Study

In this section, we conduct ablation studies for the cross-domain dataset (CD), heterogeneous domain adapter (HDA), and domain feature generator (DFG) to demonstrate their effectiveness in our DAP-MAE. All ablation studies are conducted on object classification on ScanObjectNN OBJ-BG. Table 6 presents the ablation results.

As the left side of Tab. 6 shows, the baseline model pre-trained only on single-domain data, without CD, HDA, or DFG, achieved an accuracy of 94.15% on the downstream object classification task. When combined with CD and pre-trained on a cross-domain dataset, the accuracy of the fine-tuned model was 94.32%, which is only a slight improvement. If HDA or DFG is integrated, the model's accuracy can reach 94.66%. This indicates that HDA collaboratively enhances the model's pre-training on cross-

Table 6. Left: Ablation Study of DAP-MAE on ScanObjectNN (CD: cross-domain dataset, HDA: heterogeneous domain adapter, DFG: domain feature generator). Right top: The effect of fusion and w/o fusion in the HDA. Right bottom: Effect of using different fusion modes for varying domain combinations.

CD	HDA	DFG	OBJ_BG
×	×	×	94.15
✓	×	×	94.32
✓	✓	×	94.66
✓	×	✓	94.66
✓	✓	✓	**95.18**

	OBJ_BG
w/o Fusion	94.66
Fusion	**95.18**

Fusion Mode	$\mathbb{O}+\mathbb{F}$	$\mathbb{O}+\mathbb{F}+\mathbb{S}$
Adding	92.59	92.94
FC	94.84	93.80
MLP	93.80	**95.18**

Table 7. Left: Classification performance of DAP-MAE with different feature combinations fed into the downstream task head. Right: Comparison of models' #Params (M) and FLOPs (G).

$\boldsymbol{c}$	$\boldsymbol{d}/\boldsymbol{o}$	$\mathcal{F}$	OBJ_BG
✓	×	×	93.12
×	✓	×	93.80
×	×	✓	94.49
✓	✓	×	93.28
✓	×	✓	94.66
×	✓	✓	94.15
✓	✓	✓	**95.18**

Method	#P.(M)	FL. (G)
Point-MAE [30]	22.1	4.8
Point-M2AE [58]	15.3	3.6
PointGPT-S [4]	19.5	-
ACT [7]	22.1	4.8
Point-FEMAE [57]	27.4	-
ReCon [35] (baseline)	43.6	5.3
DAP-MAE (ours)	43.8	5.4

domain datasets, and DFG extracts domain features to guide the model's adaptation during fine-tuning for downstream tasks, both effectively improving the model's performance. When HDA and DFG are used simultaneously, the model's accuracy further increases to 95.18%, which demonstrates that, as HDA and DFG play roles at different stages of DAP-MAE, they can collectively enhance the model's performance on the downstream task.

The right side of Tab. 6 presents the ablation results of the fusion methods in HDA during the fine-tuning. As shown in the upper right of Tab. 6, if HDA does not fuse the different domain data from its three MLP and only uses data from domain $\mathbb{O}$, the object classification accuracy is 94.66%, which is 0.52% lower than that (95.18%) of the model with fusion. In contrast, the lower right of Tab. 6 shows the impact of different fusion modes in HDA on the model performance. One can observe that inappropriate fusion modes in HDA during the fine-tuning significantly reduce the classification accuracy. For instance, when data of the domain $\mathbb{F}$ and $\mathbb{S}$ are sequentially added to the data of domain $\mathbb{O}$ in HDA, the object classification accuracy drops from 94.66% to 92.59% and 92.94%, respectively. If using a fully connected layer (FC) to predict the fusion coefficients, and sequentially fusing data of another two domains in HDA, object classification accuracies are 94.84% and 93.80%, respectively. When using MLPs to predict the fusion coefficients and fusing data from all three domains, the highest accuracy reaches 95.18%. These results indicate that inappropriate fusion modes in HDA may cause the model to treat data from other domains as noise for the downstream tasks, thereby reducing their accuracy.

During the fine-tuning phase of DAP-MAE, the domain features $\boldsymbol{d}$ output by DFG, the class feature $\boldsymbol{c}$, and the point cloud feature $\mathcal{F}$ are input together into the downstream task head, guiding and enhancing its performance. The left side of Tab. 7 presents the results of different combinations of these three features on downstream task. It can be observed that the class feature $\boldsymbol{c}$, domain feature $\boldsymbol{d}$, and point cloud feature $\mathcal{F}$ can all be used independently for the downstream object classification. In the pre-training phase, the class feature $\boldsymbol{c}$ is trained in an unsupervised manner, resulting in a minimum accuracy of 93.12% when used independently to fine-tune the object classification head. The domain feature $\boldsymbol{d}$ is supervised through contrastive learning during the pre-training, achieving a classification accuracy of 93.80% when used independently to fine-tune the classification head. The point cloud feature $\mathcal{F}$ is supervised by reconstruction loss during the pre-training, which preserves complete point cloud information, and achieves an accuracy of 94.49% when used to fine-tune the classification head. In the fine-tuning, combining these features generally improves the accuracy of the downstream classification, i.e. the highest accuracy reaches 95.18% when all three features are used simultaneously.

The right side of Tab. 7 compares the number of parameters (#params) and computational complexity (FLOPs) across different methods. We use ReCon [35], a dual-branch transformer with 43.6M parameters and 5.3G FLOPs, as the baseline because it enables us to better generate domain features. However, compared to ReCon, our DAP-MAE increases the number of parameters by only 0.2M and computational complexity by 0.1G, while achieving better performance on more downstream tasks.

6. Conclusion and limitation

In this paper, we propose Domain-Adaptive Point Cloud Masked Autoencoder (DAP-MAE), a unified framework for general 3D point cloud analysis across multiple domains. With a single pre-training on a cross-domain point cloud dataset, DAP-MAE leverages a heterogeneous domain adapter and a domain feature generator to learn robust and transferable representations, enabling effective adaptation to diverse downstream tasks. Experimental results confirm the strong performance of DAP-MAE across various tasks and domains. However, the current model lacks the flexibility to incorporate new domains without retraining. In future work, we plan to explore continual learning strategies to enable domain expansion and improve cross-domain generalization, particularly in settings involving cross-modal frameworks.

7. Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant 82261138629, Guangdong-Macao Science and Technology Innovation Joint Fundation under Grant 2024A0505090003, Shenzhen Municipal Science and Technology Innovation Council under Grant JCYJ20220531101412030, and Guangdong Provincial Key Laboratory under Grant 2023B1212060076.

References

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016. 3

[2] 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. 3, 5, 7

[3] 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. 2, 3, 7

[4] 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. 6, 7, 8

[5] Zhixing Chen, Di Huang, Yunhong Wang, and Liming Chen. Fast and light manifold cnn based 3d facial expression recognition across pose variations. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 229–238, 2018. 6

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 3, 5, 7

[7] 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? In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 2, 6, 7, 8

[8] Ziqi Gao, Qiufu Li, Linlin Shen, and Junpeng Yang. 3dfacemae: Pre-training of masked autoencoder using patch-based random masking reconstruction and super-resolution for 3d face recognition. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 488–503. Springer, 2024. 2

[9] Ziqi Gao, Qiufu Li, Gui Wang, and Linlin Shen. Pointfaceformer: local and global attention based transformer for 3d point cloud face recognition. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2024. 2

[10] Ankit Goyal, Hei Law, Bowei Liu, Alejandro Newell, and Jia Deng. Revisiting point cloud shape classification with a simple and effective baseline. In *International Conference on Machine Learning*, pages 3809–3820. PMLR, 2021. 6

[11] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7:187–199, 2021. 2, 6

[12] 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. 6, 7

[13] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Mvtn: Multi-view transformation network for 3d shape recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2021. 6

[14] Xu Han, Yuan Tang, Zhaoxuan Wang, and Xianzhi Li. Mamba3d: Enhancing local features for 3d point cloud analysis via state space model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4995–5004, 2024. 6, 7

[15] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6535–6545, 2021. 7

[16] Asim Jan, Huaxiong Ding, Hongying Meng, Liming Chen, and Huibin Li. Accurate facial parts localization and deep learning for 3d facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 466–472. IEEE, 2018. 6

[17] Changyuan Jiang, Shisong Lin, Wei Chen, Feng Liu, and Linlin Shen. Pointface: Point set based feature learning for 3d face recognition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021. 2

[18] Yang Jiao, Yi Niu, Yuting Zhang, Fu Li, Chunbo Zou, and Guangming Shi. Facial attention based convolutional neural network for 2d+ 3d facial expression recognition. In *2019 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2019. 6

[19] Yang Jiao, Yi Niu, Trac D Tran, and Guangming Shi. 2d+ 3d facial expression recognition via discriminative dynamic range enhancement and multi-scale learning. *arXiv preprint arXiv:2011.08333*, 2020. 6

[20] Hebeizi Li, Hongyu Yang, and Di Huang. Drfer: Learning disentangled representations for 3d facial expression recognition. *arXiv preprint arXiv:2403.08318*, 2024. 2, 6

[21] Hebeizi Li, Hongyu Yang, and Di Huang. Progressive self-supervised representation learning for 3d facial expression recognition. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2024. 3

[22] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 2018. 2, 6

[23] Chengxing Lin, Wenju Xu, Jian Zhu, Yongwei Nie, Ruichu Cai, and Xuemiao Xu. Patchmixing masked autoencoders for 3d point cloud self-supervised learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2, 7

[24] Shisong Lin, Mengchao Bai, Feng Liu, Linlin Shen, and Yicong Zhou. Orthogonalization-guided feature fusion network for multimodal 2d+ 3d facial expression recognition. *IEEE Transactions on Multimedia*, 23:1581–1591, 2020. 6

[25] Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. In *European Conference on Computer Vision*, pages 657–675. Springer, 2022. 2, 7

[26] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. *arXiv preprint arXiv:2202.07123*, 2022. 2, 6, 7

[27] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2906–2917, 2021. 2, 7

[28] Rongrong Ni, Biao Yang, Xu Zhou, Angelo Cangelosi, and Xiaofeng Liu. Facial expression recognition through cross-modality attention fusion. *IEEE Transactions on Cognitive and Developmental Systems*, 15(1):175–185, 2022. 6

[29] Oyebade K Oyedotun, Girum Demisse, Abd El Rahman Shabayek, Djamila Aouada, and Bjorn Ottersten. Facial expression recognition via joint deep learning of rgb-depth map latent representations. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3161–3168, 2017. 6

[30] 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. 2, 6, 7, 8

[31] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, Jin Chang, K. Hoffman, J. Marques, Jaesik Min, and W. Worek. Overview of the face recognition grand challenge. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 947–954 vol. 1, 2005. 3

[32] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2, 4, 6, 7

[33] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2, 6, 7

[34] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 2, 7

[35] 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. In *International Conference on Machine Learning*, pages 28223–28243. PMLR, 2023. 2, 3, 6, 7, 8

[36] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in neural information processing systems*, 35:23192–23204, 2022. 2, 6

[37] Yongming Rao, Benlin Liu, Yi Wei, Jiwen Lu, Cho-Jui Hsieh, and Jie Zhou. Randomrooms: Unsupervised pretraining from synthetic shapes and randomized layouts for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3283–3292, 2021. 7

[38] Arman Savran, Neşe Alyüz, Hamdi Dibeklioğlu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur, and Lale Akarun. Bosphorus database for 3d face analysis. In *Biometrics and Identity Management: First European Workshop, BIOID 2008, Roskilde, Denmark, May 7-9, 2008. Revised Selected Papers 1*, pages 47–56. Springer, 2008. 3, 5

[39] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019. 2

[40] Mingzhe Sui, Zhaoqing Zhu, Feng Zhao, and Feng Wu. Ffnet-m: Feature fusion network with masks for multimodal facial expression recognition. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 6

[41] Mingzhe Sui, Hanting Li, Zhaoqing Zhu, and Feng Zhao. Afnet-m: Adaptive fusion network with masks for 2d+ 3d facial expression recognition. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 116–120. IEEE, 2023. 6

[42] 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, 5

[43] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 7

[44] 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. 7

[45] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3048–3068, 2021. 6

[46] 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. 2, 6

[47] 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. 6, 7

[48] 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. 3, 5

[49] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pretraining for 3d point cloud understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 574–591. Springer, 2020. 2, 7

[50] Hongyu Yang, Kangkang Zhu, Di Huang, Hebeizi Li, Yunhong Wang, and Liming Chen. Intensity enhancement via gan for multimodal face expression recognition. *Neurocomputing*, 454:124–134, 2021. 6

[51] Honghui Yang, Tong He, Jiaheng Liu, Hua Chen, Boxi Wu, Binbin Lin, Xiaofei He, and Wanli Ouyang. Gd-mae: generative decoder for mae pre-training on lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9403–9414, 2023. 3

[52] Junpeng Yang, Qiufu Li, and Linlin Shen. Pointsurface: Discriminative point cloud surface feature extraction for 3d face recognition. *Pattern Recognition*, 156:110858, 2024. 2

[53] Xudong Yang, Di Huang, Yunhong Wang, and Liming Chen. Automatic 3d facial expression recognition using geometric scattering representation. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2015. 6

[54] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006. 3, 5

[55] 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19313–19322, 2022. 6, 7

[56] Yaohua Zha, Rongsheng Li, Tao Dai, Jianyu Xiong, Xin Wang, and Shu-Tao Xia. Sfr: Semantic-aware feature rendering of point cloud. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 6

[57] 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6962–6970, 2024. 2, 6, 7, 8

[58] 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. 6, 7, 8

[59] 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. 6, 7

[60] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10252–10263, 2021. 7

[61] Qingkai Zhen, Di Huang, Yunhong Wang, and Liming Chen. Muscular movement model-based automatic 3d/4d facial expression recognition. *IEEE Transactions on Multimedia*, 18 (7):1438–1450, 2016. 6

[62] Xiao Zheng, Xiaoshui Huang, Guofeng Mei, Yuenan Hou, Zhaoyang Lyu, Bo Dai, Wanli Ouyang, and Yongshun Gong. Point cloud pre-training with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22935–22945, 2024. 6

[63] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 2

[64] Kangkang Zhu, Zhengyin Du, Weixin Li, Di Huang, Yunhong Wang, and Liming Chen. Discriminative attention-based convolutional neural network for 3d facial expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019. 6

[65] Zhaoqing Zhu, Mingzhe Sui, Hanting Li, and Feng Zhao. Cmanet: Curvature-aware soft mask guided attention fusion network for 2d+ 3d facial expression recognition. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022. 6