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

Supplementary Material

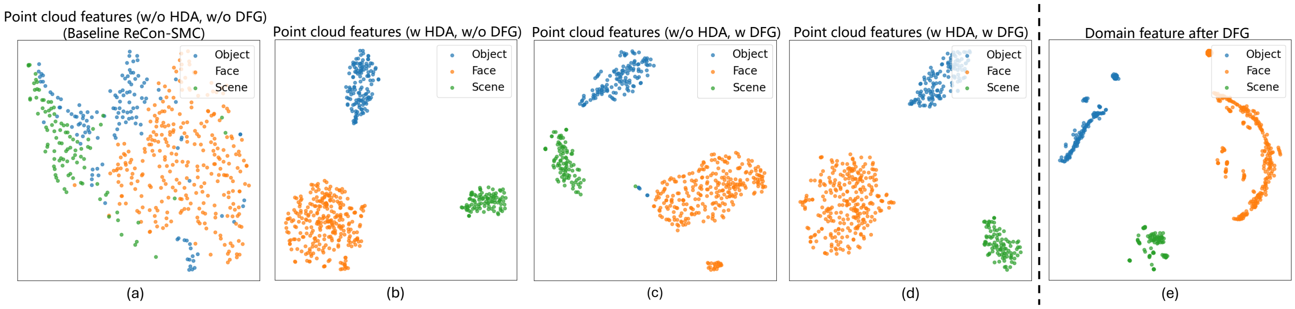

Figure 1. t-SNE visualization of features extracted from three domain datasets: (a) Point cloud features extracted w/o HDA and w/o DFG; (b) Point cloud features extracted w/ HDA and w/o DFG; (c) Point cloud features extracted w/o HDA and DFG; (d) Point cloud features extracted w/ HDA and w/ Domain DFG; (e) Domain features generated after DFG.

.1. Visualization

As shown in Fig. 1, to better evaluate how DAP-MAE can collaboratively leverage cross-domain data to enhance the feature adaptability of the model and improve the performance of downstream tasks, we designed a t-SNE [45] visualization experiment to assess the learned representations, which visualize the features extracted from the transformer encoder.

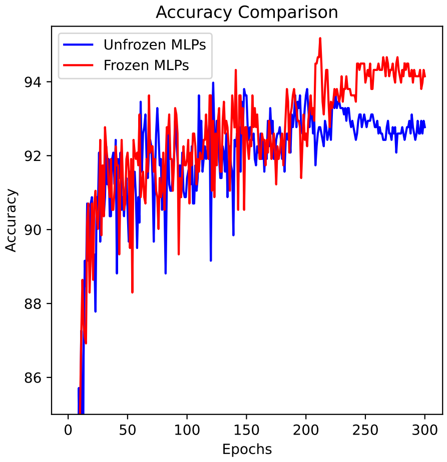

Figure 2. Comparison of classification accuracy during fine-tuning, showing how freezing or unfreezing the parameter of MLPs in HDA affects performance.

As shown in Fig. 1(a), we first present the feature distribution of our baseline model, ReCon-SMC [36], without applying the heterogeneous domain adapter (HDA) or integrating the domain feature generator (DFG), which is equivalent to the version pre-trained with simple combination of different domain data. The results show that the baseline has poor adaptability to the three different domains, as their features become intermingled, preventing the model from effectively learning and distinguishing each domain's knowledge. This confusion can even mislead the model as noise, ultimately causing a drop in performance. Fig. 1(b) presents the features precessed with HDA. We can observe that features from different domains are well separated, with tight clustering within each domain. HDA processes data from different domains using separate MLP, creating distinct feature spaces and independently learning the point cloud geometry information for each domain. This indicates that the features no longer share the same feature space, allowing better adaptation for downstream tasks without interference from other domains. Furthermore, Fig. 1(c) illustrates the features without HDA processing but concatenated with the domain feature generated by DFG. While the clustering performance is improved, some outliers still deviate from their domain centers. This situation may occur because the domain features learned by DFG come from the overall domain characteristics decomposed from point cloud features, which may not generalize well to individual samples. To further investigate, we independently visualize the domain feature in Fig. 1(e), where similar issues of domain features deviating from their cluster centers are observed. However, we can see that each domain maintains its own distinctive distribution, indicating the model's ability to learn unique domain feature patterns. When fine-tuned with tasks in the same domain, the model can leverage these patterns for rapid adaptation, leading to better performance. Finally, by combining the features processed through HDA and DFG, as shown in Fig. 1(d), the clustering is significantly improved, demonstrating the effectiveness of our two contributions.

.2. Additional experiments

Figure 2 compares classification accuracy during fine-tuning, highlighting the effects of freezing or not freezing the parameter of MLPs in HDA across fine-tuning epochs. Notably, in the final 100 epochs, the frozen MLPs approach achieves superior performance, while the unfrozen MLPs approach is prone to overfitting, resulting in a noticeable drop in accuracy.

The left side of Tab. 1 illustrates the impact of loss function weights on the performance of downstream tasks. Our total loss function is defined as:

$$\mathcal{L} = w_1 \mathcal{L}_{\text{rec}} + w_2 \mathcal{L}_{\text{con}}, \tag{1}$$

which consists of a reconstruction loss $\mathcal{L}_{\text{rec}}$ and a contrastive loss $\mathcal{L}_{\text{con}}$, balanced by weights w_1 and w_2 respectively. We can observe that increasing the weight of the reconstruction loss in the supervised setting while reducing the weight of the contrastive loss leads to better performance. This may be because the contrastive loss is prone to overfitting during pre-training.

Table 1. Comparison of loss weight settings and learning rate effects.

w_1	w_2	Accuracy	Learning Rate	Accuracy
1.0	1.0	93.80	0.001	94.84
10	0.1	94.32	0.0005	**95.18**
100	0.001	**95.18**	0.0001	94.84

The effectiveness of cross-domain data. Figure 3 shows the object classification results of models pre-trained on point clouds from one, two, and three domains, with and without DAP-MAE. One can observe that simply increasing the data does not improve and may even decrease the object classification performance. In contrast, with DAP-MAE, as the training data gradually increases, the performance also gradually improves, showing no signs of saturation.

MLP coefficients. Figure 4 shows the evolution of the coefficients generated by the MLP$^{(1)}$ and MLP$^{(2)}$ during fine-tuning on the expression recognition and object classification. The coefficients all exhibit a trend of increasing followed by decreasing, indicating that in the early stages of fine-tuning, the learning capability of HDA on point clouds from other domains is more involved in the fine-tuning. As performance on the current task domain improves, they gradually withdraw.

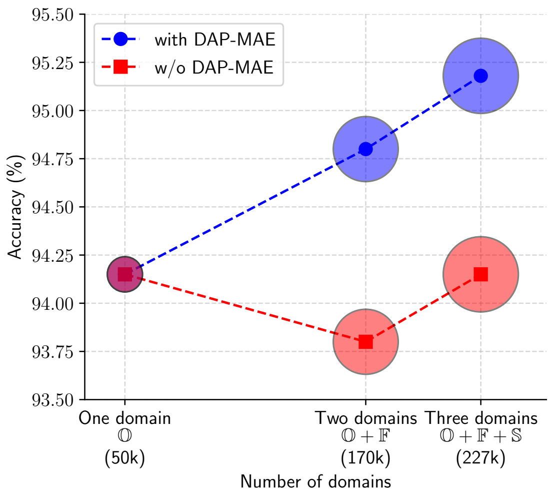

Figure 3. Effectiveness of cross-domain data.

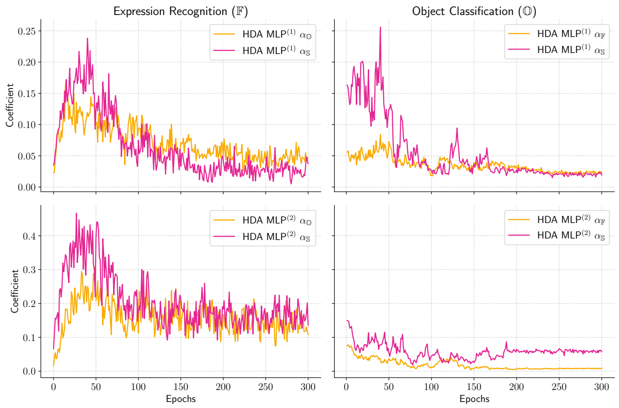

Figure 4. Coefficients learned by MLP in HDA.

.3. Experimental details.

Cross-domain dataset. ShapeNet [2] was captured from object ($\mathbb{O}$) domain, which contains more than 50,000 3D point clouds across 55 object categories. For the face domain ($\mathbb{F}$), the original FRGCv2 [32] consists of 4,007 high-quality 3D face scans from 466 individuals with expression variations. In the pre-training, we utilized the enriched FRGCv2 [10], which contains about 120K 3D faces from 1K individuals. S3DIS [1] consists of six large-scale indoor scenes from three different buildings, covering a total of 273 million points across 13 categories. Only the training split of S3DIS was used.

Fine-tuning datasets. The pre-trained DAP-MAE was fine-tuned on five datasets, each corresponding to a different downstream task: 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 object classification ($\mathbb{O}$), DAP-MAE was fine-tuned on ScanObjectNN [44], which consists of approximately 15,000 real-world objects across 15 diverse categories, and then evaluated using three different protocols, OBJ-BG,

Table 2. Training details for different downstream tasks.

Configuration	Object classification	Few-shot learning	Part segmentation	Facical expression recognition	Object detection
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
Learning rate	5e-5	5e-4	8e-5	1e-4	5e-5
Batch size	32	32	64	32	8
Weight decay	0.05	0.05	0.05	0.05	0.1
Training epochs	300	150	300	300	1080
Warm-up epochs	10	10	10	0	10
Learning rate scheduler	Cosine	Cosine	Cosine	Cosine	Cosine
Drop path rate	0.1	0.1	0.2	0.1	0.1
Number of points	2048	1024	2048	2048	40000
Number of point patches	128	64	128	128	2048
Point patch sizes	32	32	32	32	64

OBJ-ONLY, and PB-T50-RS. The few-shot learning ($\mathbb{O}$) experiments were conducted on the ModelNet40 [51] dataset, following the protocol established by [31, 40]. The experiments were structured as "n-way, m-shot", i.e. the training set contains n selected categories and m samples for each category $n \in \{5, 10\}$ and $m \in \{10, 20\}$. Part segmentation ($\mathbb{O}$) was conducted on ShapeNetPart [2] which consists of 16,881 objects spanning 16 categories.

For facial expression recognition ($\mathbb{F}$), DAP-MAE was respectively fine-tuned on BU-3DFE [57] and Bosphorus [39]. BU-3DFE contains 2,500 scans of 100 individuals (56 females and 44 males) aged between 18 and 70. Each individual has 25 samples representing seven different expressions: one neutral expression and six basic expressions with four different intensities. Bosphorus contains a total of 4,666 3D face scans collected from 105 individuals aged between 25 and 35. Among them, 65 individuals exhibit the six basic expressions with single intensity.

For 3D object detection ($\mathbb{S}$), DAP-MAE was evaluated on ScanNetV2 [6], which consists of real-world richly annotated 3D point clouds of indoor scenes. ScanNetV2 comprises 1201 training scenes, 312 validation scenes, and 100 hidden test scenes. In ScanNetV2, 18 object categories are labeled using axis-aligned bounding boxes.

Experiment setting. As Tab. 2 shows, the batch size was set to 512, and DAP-MAE was optimized using the AdamW optimizer with an initial learning rate of 0.0005 and a weight decay of 0.05. While the learning rate was decayed by a cosine schedule with a warm-up period of 10 epochs, the total epoch number was 300. Random scaling and translation were used for data augmentation. Also Tab. 2 shows the fine-tune details on various downstream tasks. All experiments were conducted on an NVIDIA V100 GPU (32GB).