Reconstruction Target Matters in Masked Image Modeling for Cross-Domain Few-Shot Learning

Ran Ma, Yixiong Zou*, Yuhua Li, Ruixuan Li

School of Computer Science and Technology, Huazhong University of Science and Technology
{ranma,yixiongz,idcliyuhua,rxli}@hust.edu.cn

Abstract

Cross-Domain Few-Shot Learning (CDFSL) requires the model to transfer knowledge from the data-abundant source domain to data-scarce target domains for fast adaptation, where the large domain gap makes CDFSL a challenging problem. Masked Autoencoder (MAE) excels in effectively using unlabeled data and learning image's global structures, enhancing model generalization and robustness. However, in the CDFSL task with significant domain shifts, we find MAE even shows lower performance than the baseline supervised models. In this paper, we first delve into this phenomenon for an interpretation. We find that MAE tends to focus on low-level domain information during reconstructing pixels while changing the reconstruction target to token features could mitigate this problem. However, not all features are beneficial, as we then find reconstructing high-level features can hardly improve the model's transferability, indicating a trade-off between filtering domain information and preserving the image's global structure. In all, the reconstruction target matters for the CDFSL task. Based on the above findings and interpretations, we further propose **D**omain-**A**gnostic **M**asked **I**mage **M**odeling (DAMIM) for the CDFSL task. DAMIM includes an Aggregated Feature Reconstruction module to automatically aggregate features for reconstruction, with balanced learning of domain-agnostic information and images' global structure, and a Lightweight Decoder module to further benefit the encoder's generalizability. Experiments on four CDFSL datasets demonstrate that our method achieves state-of-the-art performance.

Introduction

With the advancement of deep learning, neural networks have become increasingly capable of handling large datasets. However, collecting large amounts of data is difficult in some cases, such as healthcare, leading to the emergence of Cross-Domain Few-Shot Learning (CDFSL). This task involves first training a model on a large-scale dataset (source domain, such as a general dataset like ImageNet (Deng et al. 2009)) and then transferring it to specialized downstream datasets (target domain, such as medical datasets (Codella et al. 2018)) where only a few samples are available. Significant discrepancies between the source and target datasets usually make the transfer and downstream learning difficult (Chen et al. 2018; Guo et al. 2020). Therefore, addressing domain gaps is crucial for the CDFSL task.

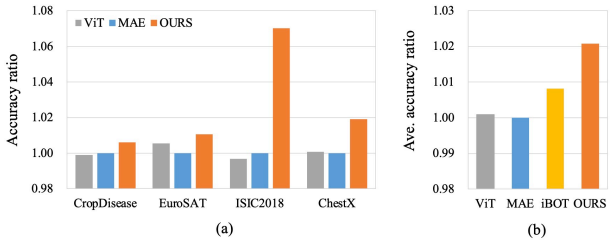

Figure 1: (a) The ratio of the accuracy of ViT, MAE, and our method on four CDFSL datasets, where we can see MAE underperforms on these datasets. (b) The average performance of ViT, MAE, iBOT, and our method, inspires us to think about the role of reconstruction target in MIM for CDFSL.

Masked Autoencoder (MAE) (He et al. 2022) is a self-supervised learning method that focuses on predicting the masked parts of the image, thereby deepening the model's understanding of the image's global information. This approach has shown strong generalization capabilities across a wide range of downstream tasks (He et al. 2022; Kong et al. 2023; Zhang et al. 2024). However, when applying MAE to train models for CDFSL tasks, we find that its performance[1] in the target domain is often unsatisfactory (Fig. 1a), even lagging behind that of the Vision Transformers (ViT) (Dosovitskiy et al. 2021) trained with supervised learning. This counter-intuitive phenomenon makes us wonder *what is it that makes MAE less effective under large domain gaps*.

In this paper, we first delve into this phenomenon for an interpretation. We first consider Masked Image Modeling (Bao et al. 2022) (MIM), the superset of MAE, to see whether other methods exhibit similar behavior. We find that iBOT (Zhou et al. 2022) outperforms MAE under large domain gaps (Fig. 1b). Since it takes token features as the reconstruction target while MAE takes image pixels, it raises questions about the role of pixel-level reconstruction. Subsequently, we find MAE tends to absorb domain information during pixel reconstruction, represented as low-level pixel

*Corresponding author.

[1] As the accuracy of different datasets varies largely (e.g., 90 vs. 20), we use the ratio of accuracies against MAE for illustration.

details. In contrast, shifting the reconstruction target to token features mitigates this problem, suggesting that domain information is filtered out in token features. However, not all features are beneficial. We find reconstructing high-level features cannot benefit transferability, indicating a trade-off between filtering out low-level domain information and preserving the image's global information. In all, the reconstruction target matters in MIM for the CDFSL task.

Based on the above findings and interpretations, we further propose a novel approach called **D**omain-**A**gnostic **M**asked **I**mage **M**odeling (DAMIM) for the CDFSL task. DAMIM includes two main components: the Aggregated Feature Reconstruction (AFR) module and a Lightweight Decoder (LD) module. The AFR module generates effective domain-agnostic reconstruction targets, reducing the learning of domain-specific information that impedes generalization across different domains, while preserving the image's global information. Complementing the AFR module, since the reconstruction target is simplified from pixels to features, the LD module is introduced to prevent the encoder from relying heavily on the decoder, further enhancing the feature encoder's generalization. Experimental analysis validates DAMIM's advantages, with results demonstrating that it significantly improves MIM's generalization and achieves outstanding performance on CDFSL datasets.

To summarize, our contributions are as follows:

- To the best of our knowledge, we are the first to reveal the limited performance of pixel-based MIM in CDFSL.
- Through analysis, we find that the limited performance is due to the tendency to learn low-level domain information during pixel reconstruction and point out that the reconstruction target matters in MIM for CDFSL.
- We propose Domain-Agnostic Masked Image Modeling, a novel approach that includes an Aggregated Feature Reconstruction module to automatically aggregate features for reconstruction, balancing learning of domain-agnostic information and image's global structure, and a Lightweight Decoder module to further benefit the encoder's generalizability.
- Extensive experiments validate our analysis and methodology, demonstrating that our approach achieves state-of-the-art performance on CDFSL datasets.

Related Work

Cross-Domain Few-Shot Learning

CDFSL aims to transfer knowledge from a well-trained source domain to a different target domain with limited labeled data. CDFSL is mainly studied through two approaches: transferring-based (Guo et al. 2020; Phoo and Hariharan 2021; Hu and Ma 2022) methods, which adapt pre-trained models from large-scale source datasets to target domains with limited data, and meta-learning (Tseng et al. 2020; Wang and Deng 2021), which focuses on training models to quickly adapt to new tasks. In contrast to these methods, our approach emphasizes source-domain training while simultaneously enhancing knowledge transfer and target-domain fine-tuning.

Masked Image Modeling

MIM is a self-supervised learning method that trains models to reconstruct masked parts of an image, promoting learning valuable image representations. MIM research can be categorized by reconstruction target: pixel-based (He et al. 2022; Liu et al. 2023; Xie et al. 2022) and token-based (Zhou et al. 2022; Chen et al. 2022; Bao et al. 2022) methods. Pixel-based MIM methods, such as the MAE (He et al. 2022), focus on reconstructing the missing pixels from the surrounding visible areas. Token-based MIM methods, in contrast, predict higher-level representations or tokens derived from the image. These methods leverage the semantic information encoded in tokens, which has shown to be highly effective in various visual recognition tasks. Despite these advances, it is important to note that no studies have specifically investigated the performance of MAE in CDFSL tasks, leaving an open area for future research.

Delve into MAE on CDFSL

In this section, we explore the reasons why MAE underperforms on CDFSL tasks with large domain gaps.

Preliminaries

Cross-Domain Few-Shot Learning (CDFSL) aims to adapt a model trained on a source domain $\mathcal{D}^S = \{(x_i^S, y_i^S)\}_{i=1}^{N_S}$ with large, diverse samples to a target domain $\mathcal{D}^T = \{(x_i^T, y_i^T)\}_{i=1}^{N_T}$, which has only a few labeled samples N_T and a significant domain gap from the source domain. During adaption and evaluation on $\mathcal{D}^T$, existing research (Snell, Swersky, and Zemel 2017; Guo et al. 2020) employ a k-way n-shot paradigm to sample from $\mathcal{D}^T$, forming small datasets (episodes) containing k classes with n training samples each. The model is trained on these $k \cdot n$ samples (support set, $\{x_{ij}, y_i\}_{i=1,j=1}^{k,n}$) and then tested on k classes (query set, $\{x^q\}$). This approach, alongside addressing the significant domain gap, ensures the model generalizes well to new, unseen samples in the target domain.

Masked Autoencoder (MAE), as in Fig. 2, uses the architecture of an autoencoder, training the model by masking parts of the image and requiring the model to reconstruct the masked regions. Given an input image $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, following ViT (Dosovitskiy et al. 2021), MAE first divides the image into regular, non-overlapping patches:

$$\mathbf{X} = \{\mathbf{X}_i\}_{i=1}^{N}, \quad \mathbf{X}_i \in \mathbb{R}^{P \times P \times C} \tag{1}$$

where N is the number of patches and P is the patch size. Then, with a mask ratio r, some patches are randomly masked. The mask is defined as a vector $\mathbf{m} \in \{0,1\}^N$, where $N \times r$ elements are randomly set to 0, and the remaining elements are set to 1, we use $\mathbf{m}$ to mask patches:

$$\mathbf{X}^{vis} = \mathbf{X} \odot \mathbf{m}. \tag{2}$$

$\mathbf{X}^{vis} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_{N \times (1-r)}}]$ is the visible patches set, and $\mathbf{X}^{mask} = [\mathbf{x_{N \times (1-r)+1}}, \mathbf{x_{N \times (1-r)+2}}, \ldots, \mathbf{x_N}]$ denotes the masked patches. The encoder f_{enc} processes visible patches $\mathbf{X}^{vis}$ as input and outputs the representations:

$$Z = f_{enc}(\mathbf{X}^{vis}), \quad Z \in \mathbb{R}^{(N \times (1-r)) \times d} \tag{3}$$

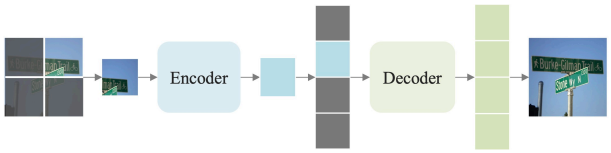

Figure 2: The structure of pixel-based MAE.

where d is the representation dimension. Then the decoder f_{dec} receives Z and the encoded mask patches $M \in \mathbb{R}^{(N\times r)\times d}$, and outputs the reconstructed image:

$$\hat{\mathbf{X}} = f_{dec}([Z, M]), \quad \hat{\mathbf{X}} = [\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_N]. \tag{4}$$

The model is trained by minimizing the difference between the reconstructed image and the original image, typically using the Mean Squared Error (MSE) as the loss function:

$$\mathcal{L} = \frac{1}{N \times r} \sum_{i \in \mathbf{X}^{mask}} \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2 \tag{5}$$

In this paper, we use unsupervised MAE to train the encoder on the source domain without requiring labels. Then we discard the decoder, retain the encoder as the backbone, and use ProtoNet (Liu, Song, and Qin 2020) with a distance function $d(\cdot, \cdot)$ for the target domain few-shot evaluation:

$$\hat{y_q} = arg\min_i d(\frac{1}{n}\sum_j f_{enc}(x_{ij}), f_{enc}(x_q)). \tag{6}$$

Delve into MAE's limited performance in CDFSL

MAE tends to learn low-level information. We find that the token-based iBOT (Zhou et al. 2022) outperforms the pixel-based MAE in CDFSL tasks, inspiring us to consider the role of pixel-level reconstruction. Specifically, we hypothesize that MAE's focus on pixel reconstruction might contribute to its ineffectiveness in CDFSL tasks, as it requires the model to learn low-level visual information, such as the brightness or color of images that could be relevant to domains. To investigate the learning behavior of MAE during reconstruction, we set different features in the ViT as the reconstruction target. As shown in Fig. 3(a), the reconstruction loss for shallow-layer features is significantly lower than that for deeper-layer ones, indicating that the model is more inclined to capture the low-level information. This tendency might lead the MAE to prioritize learning low-level information over more semantic and global representations.

Low-level information exhibits domain specificity. To delve into the properties of low-level information, we apply a masking operation to disrupt the output of different ViT layers. Then, we measure the domain similarity by the CKA[2] similarity (Kornblith et al. 2019) of the final output features between source and target domains, as shown in Fig. 3(b). When low-level information is disrupted, domain similarity increases significantly, compared to when higher-level features are disrupted. This suggests that low-level information is highly domain-specific, containing rich domain

[2]Please see the appendix for details.

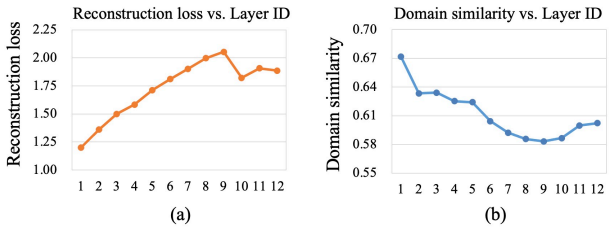

Figure 3: (a) Reconstruction loss measured using different ViT layer features as targets. Lower loss for shallow-layer features suggests the model easily captures low-level features. (b) Domain similarity of final features between source and target domains after disrupting different layers. Disrupting shallow-layer features increases domain similarity.

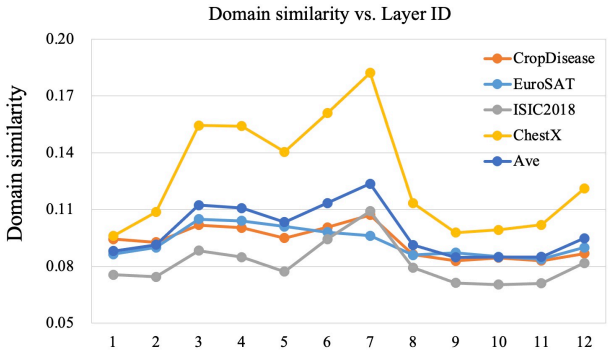

Figure 4: Domain similarity of models using different ViT layer features as reconstruction targets. Shallow-layer features have lower domain similarity, and deeper-layer features do not significantly enhance transferability, indicating a trade-off between filtering domain information and preserving the image's global structure.

information closely tied to the source domain. Therefore, we conclude that MAE tends to prioritize lower-level information during the reconstruction process, and the domain information carried by it hampers the model's ability to generalize effectively from the source domain to target domains. On the other hand, as network depth increases, such information is gradually filtered out, allowing the model to capture more semantic information than low-level information. As a result, it might improve the model's generalizability to replace pixels with token features as the reconstruction target.

Reconstruction with high-level features. Since pixel-level or low-level features are rich in domain-specific information, we are motivated to explore using higher-level features for reconstruction. As shown in Fig. 4, when different layer features are used as reconstruction targets, low-level features exhibit lower domain similarity between source and target domains, consistent with previous findings. However, not all features are beneficial to domain generalization, especially higher-level features near the network's output. In shallow layers, the model predominantly captures low-level information, such as color distribution and brightness, which are often domain-specific and difficult to transfer effectively.

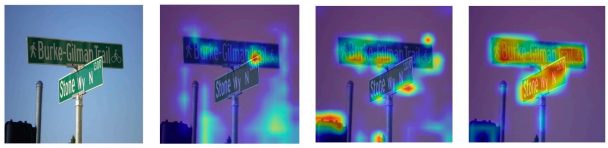

Figure 5: Visualization of MAE's features. The model captures low-level information in shallow layers and shifts focus to semantic parts as the network deepens.

As the network goes deeper, its focus gradually shifts towards more abstract and semantic information, as depicted in Fig. 5. These higher-level features tend to activate only the regions for semantic objects, with less focus on other background regions, limiting MIM's ability to learn the image's global structure. Therefore, higher-level features result in a sharp decline in domain similarity. Consequently, there is a trade-off between filtering out low-level domain-specific information and preserving the image's global information.

Conclusion and Discussion

The reconstruction target matters for MIM in the CDFSL task. For low-level features or pixels, domain information is carried as the low-level information (e.g., color distribution, brightness, etc.) that harms generalization. For high-level features, the model tends to focus on only the semantic parts, limiting MIM's learning of the image's global structure. Therefore, methods for automatically selecting the reconstruction target are necessary.

Method

In this section, we introduce Domain-Agnostic Masked Image Modeling (DAMIM) for CDFSL, as shown in Fig. 6. We propose a Feature Aggregation Reconstruction (AFR) module, which integrates features from multiple layers to serve as the reconstruction target. Complementing AFR, we design a Lightweight Decoder (LD) module that prevents the encoder from over-relying on the decoder for reconstruction, thus enhancing the feature encoder's generalization. The following sections detail each component.

Aggregated Feature Reconstruction Module To balance filtering low-level domain information while preserving global information, we propose the AFR Module, automatically aggregating the most advantageous features as a reconstruction target to achieve this balance.

Firstly, we establish an auxiliary encoder that shares weights with the original encoder to extract features from multiple layers. The original encoder processes only the visible patches of an image, while the auxiliary encoder processes all patches. Following MAE (He et al. 2022), the input image $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ is divided into non-overlapping patches $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$, where each patch $\mathbf{X}_i \in \mathbb{R}^{P\times P\times C}$. A high proportion of these patches is then randomly masked, resulting in a visible patch set $\mathbf{X}^{vis} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N'}]$ and a masked patch set $\mathbf{X}^{mask} = [\mathbf{x}_{N'+1}, \mathbf{x}_{N'+2}, \ldots, \mathbf{x}_N]$, where N' is the number of visible patches. The original encoder processes the visible patches to obtain their latent representations:

$$\mathcal{Z} = \mathbf{Encoder}(\mathbf{X}^{vis}), \quad \mathcal{Z} \in \mathbb{R}^{N'\times d}. \tag{7}$$

Simultaneously, the auxiliary encoder processes full image patches and outputs features at the l-th layer:

$$f^{(l)} = \mathbf{Encoder}^{(l)}_{aux}(\mathbf{X}), \quad f^{(l)} \in \mathbb{R}^{N\times d}. \tag{8}$$

$f^{(l)}$ represents the output of the l-th layer of the auxiliary encoder $\mathbf{Encoder}^{(l)}_{aux}$. Then, we introduce a feature aggregation module to fuse layer-wise features, and a projection layer aligns the feature spaces across different levels. For a given layer l, the feature alignment is achieved by:

$$\tilde{f}^{(l)} = W^{(l)} f^{(l)}, \quad W^{(l)} \in \mathbb{R}^{d\times d} \tag{9}$$

where $W^{(l)}$ is the projection matrix for the l-th layer. To combine these aligned features, we use a weighted average pooling mechanism defined as:

$$\mathcal{F} = \sum_{l=1}^{L} \alpha^{(l)} \tilde{f}^{(l)}, \mathcal{F} \in \mathbb{R}^{N\times d}. \tag{10}$$

where $\alpha^{(l)}$ are feature weights, which are generated based on the layer reconstruction loss with a linear and softmax layer. The aggregated feature $\mathcal{F}$ serves as the reconstruction target, enabling effective feature combination across layers.

Lightweight Decoder Since the reconstruction target is feature-based, we design a lightweight decoder to prevent the encoder from over-relying on the decoder during the MAE reconstruction process, ensuring the encoder focuses on learning more generalizable features.

Our decoder consists of a single Transformer block with just one attention head, significantly reducing computational complexity. Instead of the conventional query-key attention mechanism (Vaswani et al. 2017), we use a simple cosine similarity score between tokens, as shown in Fig. 7:

$$\text{CosSim}_{ij} = \frac{\mathbf{t}_i \cdot \mathbf{t}_j}{\|\mathbf{t}_i\|\|\mathbf{t}_j\|}, \quad i, j = 1, \ldots, N \tag{11}$$

where $\mathbf{t}_i$ denotes an image token. This substitution not only simplifies the architecture but also better aligns with our goal of reducing the decoder's influence. We also remove the MLP (Multi-Layer Perceptron) module entirely, further streamlining the decoder. We then concatenate the visible representations $\mathcal{Z}$, processed by the encoder, with the mask tokens $M \in \mathbb{R}^{(N-N')\times d}$, which are obtained by embedding the mask patches and adding positional information:

$$\mathcal{R} = \mathbf{Decoder}(\mathcal{Z}, M), \quad \mathcal{R} \in \mathbb{R}^{n\times d} \tag{12}$$

The processed output from our decoder yields the final reconstruction $\mathcal{R}$.

Reconstruction Loss The final reconstruction loss $\mathcal{L}_{recon}$ is calculated using the Mean Squared Error (MSE) between the reconstructed output $\mathcal{R}$ and the target aggregated features $\mathcal{F}$, which is defined as:

$$\mathcal{L}_{recon} = \frac{1}{N} \sum_{i=1}^{N} \|\mathcal{F}_i - \mathcal{R}_i\|^2 \tag{13}$$

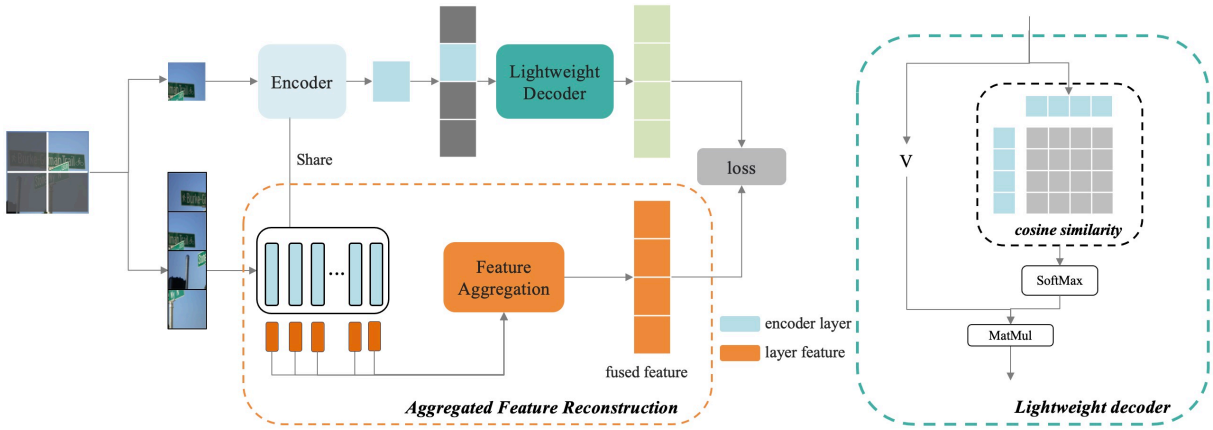

Figure 6: Our proposed **DAMIM** model comprises two key components: (a) The Aggregated Feature Reconstruction (AFR) module (in the orange box), which integrates multi-layer features as reconstruction targets. (b) The Lightweight Decoder (LN) module (in the green box) substitutes the original query-key attention in ViT with a cosine similarity in a simplified structure.

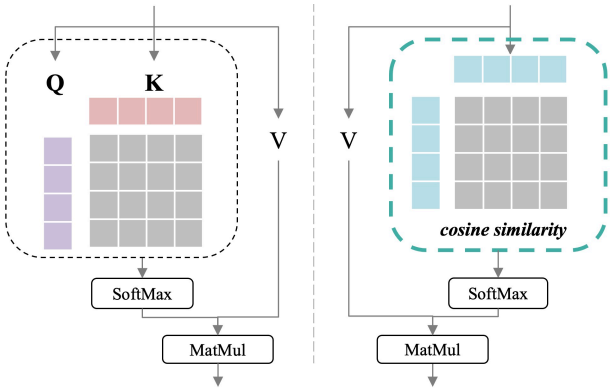

Figure 7: Our Lightweight Decoder. LD uses cosine similarity between tokens instead of a query-key mechanism, reducing computation while capturing essential relationships.

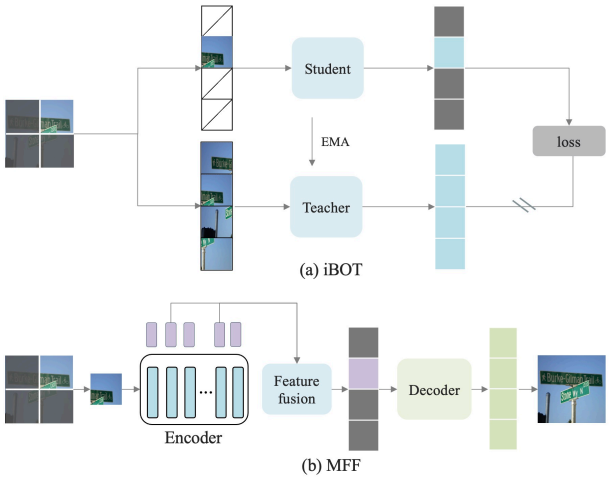

Figure 8: Structures of iBOT and MFF. iBOT uses a teacher-student framework with mask tokens visible in the encoder, while DAMIM employs a lightweight decoder. MFF fuses features before decoding, while DAMIM uses aggregated features for reconstruction.

Target domain evaluation and fine-tuning During the evaluation and fine-tuning phase in the target domain, the decoder is discarded, and the encoder is retained. Following recent works (Hu et al. 2022; Fu et al. 2023), we perform prototype-based classification within few-shot episodes, as described in Eq.(6), or fine-tune the backbone using the support set for the classifier-based classification.

Experiments

Dataset

Following the BSCD benchmark (Guo et al. 2020), we use miniImagenet (Vinyals et al. 2016) as the source dataset and four cross-domain datasets as target datasets: CropDisease (Mohanty, Hughes, and Salathé 2016), EuroSAT(Helber et al. 2019), ISIC2018 (Codella et al. 2018) and ChestX (Wang et al. 2017). MiniImageNet has 100 natural image categories with 600 images each, split into 64 training, 16 validation, and 20 test categories. CropDisease includes 38 crop disease categories with 43,456 images. EuroSAT is used for land use and cover classification from satellite images, with 27,000 images across 10 categories. ISIC2018 includes 10,015 skin lesion images in 7 categories. ChestX focuses on thoracic disease diagnosis, with 25,847 images across 7 categories.

Implementation Details

Follow StyleAdv (Fu et al. 2023), we use ViT-S as the backbone, initialized with DINO (Caron et al. 2021) pretraining on ImageNet1K (Deng et al. 2009). During base class training, we train the model with AdamW (Loshchilov and Hutter 2019) with a learning rate of 0.001 for the classifier, 1e-7 for the backbone, and 1e-6 for the decoder. For novel-class fine-tuning, we discard the decoder and fine-tune the backbone using the SGD optimizer with a momentum of 0.9.

Method	Mark	FT	ChestX	ISIC2018	EuroSAT	CropDisease	Average
MEM-FS (Walsh, Osman, and Shehata 2023)	TIP-23	×	22.76	32.97	68.11	81.11	51.24
StyleAdv (Fu et al. 2023)	CVPR-23	×	22.92	33.05	72.15	81.22	52.34
SDT (Liu et al. 2024)	NN-24	×	22.79	33.40	72.71	81.03	52.48
FLoR (Zou et al. 2024)	CVPR-24	×	22.78	34.20	72.39	81.81	52.80
DAMIM	**Ours**	×	**22.97**	**34.66**	**72.87**	**82.34**	**53.21**
PMF (Hu et al. 2022)	CVPR-22	✓	21.73	30.36	70.74	80.79	50.91
FLoR (Zou et al. 2024)	CVPR-24	✓	23.26	35.49	73.09	83.55	53.85
StyleAdv (Fu et al. 2023)	CVPR-23	✓	22.92	33.99	**74.93**	**84.11**	53.99
DAMIM	**Ours**	✓	**23.38**	**36.35**	73.61	83.90	**54.31**
MEM-FS+RDA* (Walsh, Osman, and Shehata 2023)	TIP-23	✓	23.85	37.07	75.91	83.74	55.14
DAMIM*	**Ours**	✓	**23.91**	**38.07**	**77.23**	**86.74**	**56.49**

Table 1: Comparison with the state-of-the-art works based on ViT-S by 5-way 1-shot accuracy.

Method	Mark	FT	ChestX	ISIC2018	EuroSAT	CropDisease	Average
MEM-FS (Walsh, Osman, and Shehata 2023)	TIP-23	×	26.67	47.38	86.49	93.74	63.57
StyleAdv (Fu et al. 2023)	CVPR-23	×	26.97	47.73	88.57	94.85	64.53
SDT (Liu et al. 2024)	NN-24	×	26.72	47.64	89.60	95.00	64.75
FLoR (Zou et al. 2024)	CVPR-24	×	26.71	49.52	**90.41**	95.28	65.48
DAMIM	**Ours**	×	**27.28**	**50.76**	89.50	**95.52**	**65.77**
PMF (Hu et al. 2022)	CVPR-22	✓	27.27	50.12	85.98	92.96	64.08
StyleAdv (Fu et al. 2023)	CVPR-23	✓	26.97	51.23	90.12	95.99	66.08
FLoR (Zou et al. 2024)	CVPR-24	✓	27.02	53.06	90.75	**96.47**	66.83
DAMIM	**Ours**	✓	**27.82**	**54.86**	**91.18**	96.34	**67.55**
MEM-FS+RDA* (Walsh, Osman, and Shehata 2023)	TIP-23	✓	27.98	51.02	88.77	95.04	65.70
DAMIM*	**Ours**	✓	**28.10**	**55.44**	**91.08**	**96.49**	**67.78**

Table 2: Comparison with the state-of-the-art works based on ViT-S by 5-way 5-shot accuracy.

AFR	LD	CropDisease	EuroSAT	ISIC2018	ChestX	Ave.
		94.88	86.94	46.00	26.48	63.58
✓		95.25	88.63	49.87	26.58	65.08
	✓	95.21	88.90	49.74	27.10	65.24
✓	✓	**95.52**	**89.50**	**50.38**	**27.28**	**65.77**

Table 3: Ablation study by the 5-way 5-shot accuracy.

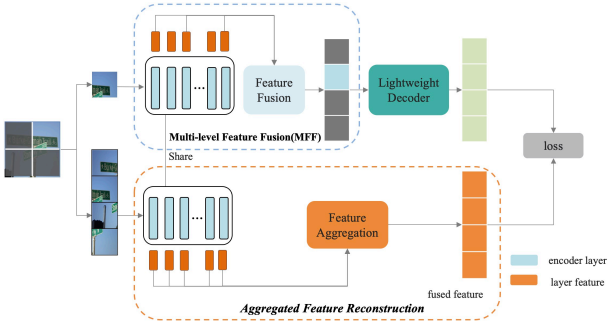

Figure 9: Applying our method to MFF. MFF improves upon the pixel-based MAE by fusing features from multiple encoder blocks before the decoder. Our approach explicitly avoids learning domain-specific features by using aggregated multi-layer features as the reconstruction target.

Comparison with SOTA Method

Tab. 1 and Tab. 2 compare our results with state-of-the-art methods using the ViT-S backbone pretrained by DINO in 1-shot and 5-shot settings. We distinguish results with and without fine-tuning (FT), with an asterisk (*) indicating a transductive setting. Comparisons with CNN backbone methods are provided in the supplementary materials. PMF (Hu et al. 2022), StyleAdv (Fu et al. 2023), MEM-FS (Walsh, Osman, and Shehata 2023), SDT (Liu et al. 2024) and FLoR (Zou et al. 2024) are introduced as competitors. Our consistently superior performance across all settings highlights the effectiveness of our DAMIM.

Comparison with other MIM Methods

We compare DAMIM with the token-based iBOT (Zhou et al. 2022) and the pixel-based MFF (Liu et al. 2023), as shown in Tab. 4. The structures of these two models are depicted in Fig. 8. iBOT uses a teacher-student framework, where the student reconstructs masked patches using the teacher's output. The mask tokens are visible to the entire encoder in iBOT, while DAMIM uses a lightweight decoder for reconstruction. MFF enhances MAE by fusing multiple encoder block features before decoding, implicitly prevent-

Method	CropDisease	EuroSAT	ISIC2018	ChestX	Ave.
iBOT	94.94	**89.58**	48.34	26.97	64.96
MFF	95.15	89.16	49.83	27.22	65.34
DAMIM	**95.52**	89.50	**50.76**	**27.28**	**65.77**
iBOT	94.94	89.58	48.34	26.97	64.96
iBOT+DAMIM	**95.30**	89.57	**48.67**	**27.02**	**65.14**
MFF	95.15	89.16	49.83	27.22	65.34
MFF+DAMIM	**95.28**	**89.51**	**51.21**	**27.85**	**65.96**

Table 4: Comparison with iBOT and MFF by 5-way 5-shot accuracy. DAMIM outperforms both methods.

Method	CropDisease	EuroSAT	ISIC2018	ChestX	Ave.
BL	0.0982	0.0907	0.0705	0.0858	0.0863
BL+AFR	0.1071	0.0961	0.1091	0.1822	0.1236
BL+AFR+LD	0.1305	0.1226	0.1254	0.2252	0.1509

Table 5: Both AFR and LD improve domain similarity.

ing overfitting to low-level features. In contrast, DAMIM explicitly avoids learning low-level domain information by using aggregated multilayer features as the reconstruction target. The results show that DAMIM outperforms both iBOT and MFF, and applying our method to iBOT and MFF further boosts performance. Notably, combining our approach with MFF, as shown in Fig. 9, achieves a new state-of-the-art, highlighting its effectiveness and adaptability.

Ablation Study

Vertification of DAMIM We conduct an ablation study on DAMIM, as shown in Tab. 3. The results demonstrate that both the Aggregated Feature Reconstruction module and the Lightweight Decoder module independently improve the baseline by 1.50% and 1.66%, respectively. Their combination further boosts performance by 2.19%, highlighting these components' critical contributions to our method.

Verification of AFR The AFR module aims to balance filtering low-level domain information while preserving global structure. To evaluate this, We measure domain similarity using CKA similarity (Kornblith et al. 2019; Li, Liu, and Bilen 2021) between the source and target domains, comparing features before and after applying AFR. As shown in Tab. 5, AFR enhances domain similarity, indicating that the encoder's features effectively filter out domain-specific information while retaining more generalizable content.

Verification of LD The LD module reduces the encoder's reliance on the decoder for reconstruction, encouraging generalization and effective learning. As shown in Tab. 5, combining it with AFR further improves domain similarity, indicating better generalization. We also ablate our design of the LD module in Tab. 6. It is evident that a lightweight decoder, which removes the MLP module and uses a cosine similarity map between tokens instead of the query-key mechanism, achieves superior performance compared to other designs.

light	MLP	correlation	CropDisease	EuroSAT	ISIC2018	ChestX	Ave.
	✓	Attn.	94.94	88.56	47.43	26.77	64.43
✓	✓	Attn.	94.94	88.26	49.47	27.02	64.93
✓	✓	Iden.	94.87	88.92	47.60	26.80	64.55
✓	✓	Euc.	94.72	88.18	50.06	27.09	65.01
✓	✓	Cos.	94.70	88.10	**50.35**	27.10	65.06
✓		Iden.	94.81	88.99	47.59	26.79	64.55
✓		Euc.	95.01	**89.08**	49.55	26.96	65.15
✓		Cos.	**95.21**	88.90	49.74	**27.10**	**65.24**

Table 6: Ablation study of LD by 5-way 5-shot accuracy.

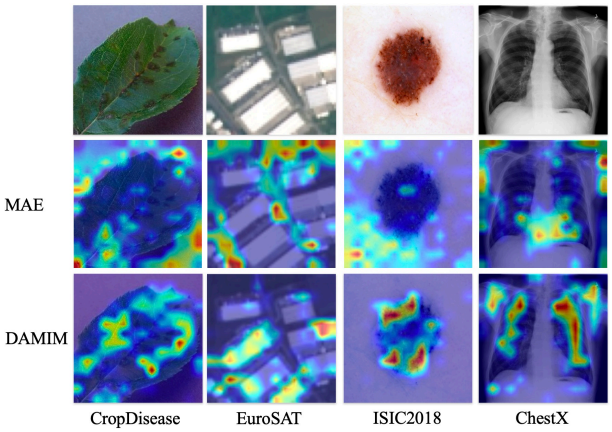

Figure 10: Heatmaps for MAE and DAMIM in four target domains. DAMIM captures meaningful regions and comprehensive object information more effectively.

Visualization

We generate heatmaps for MAE and DAMIM using Gram-CAM (Selvaraju et al. 2017). As shown in Fig. 10, DAMIM effectively captures meaningful regions and comprehensive object information, indicating better generalization of knowledge from the source to target domains.

Conclusion

In this paper, We find that MAE focuses on low-level domain information during pixel reconstruction, while high-level feature reconstruction struggles with transferability, indicating a trade-off between filtering domain information and preserving global structure. We propose DAMIM, including an Aggregated Feature Reconstruction module for balanced learning and a Lightweight Decoder module to further boost generalization. Experiments show that DAMIM outperforms state-of-the-art methods.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grants 62206102, 62436003, 62376103 and 62302184; the Science and Technology Support Program of Hubei Province under grant 2022BAA046; Hubei science and technology talent service project under grant 2024DJC078.

References

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2022. BEiT: BERT Pre-Training of Image Transformers. In *International Conference on Learning Representations*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Chen, W.-Y.; Liu, Y.-C.; Kira, Z.; Wang, Y.-C. F.; and Huang, J.-B. 2018. A closer look at few-shot classification. In *Proceedings of the International Conference on Learning Representations*.

Chen, Y.; Liu, Y.; Jiang, D.; Zhang, X.; Dai, W.; Xiong, H.; and Tian, Q. 2022. SdAE: Self-distillated Masked Autoencoder. In Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision – ECCV 2022*, 108–124. Cham: Springer Nature Switzerland.

Codella, N. C. F.; Gutman, D.; Celebi, M. E.; Helba, B.; Marchetti, M. A.; Dusza, S. W.; Kalloo, A.; Liopyris, K.; Mishra, N.; Kittler, H.; and Halpern, A. 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 168–172.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Fu, Y.; Xie, Y.; Fu, Y.; and Jiang, Y.-G. 2023. StyleAdv: Meta Style Adversarial Training for Cross-Domain Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24575–24584.

Guo, Y.; Codella, N. C.; Karlinsky, L.; Codella, J. V.; Smith, J. R.; Saenko, K.; Rosing, T.; and Feris, R. 2020. A broader study of cross-domain few-shot learning. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, 124–141. Springer.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000–16009.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226.

Hu, S. X.; Li, D.; Stühmer, J.; Kim, M.; and Hospedales, T. M. 2022. Pushing the Limits of Simple Pipelines for Few-Shot Learning: External Data and Fine-Tuning Make a Difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9068–9077.

Hu, Y.; and Ma, A. J. 2022. Adversarial Feature Augmentation for Cross-domain Few-Shot Classification. In Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision – ECCV 2022*, 20–37. Cham: Springer Nature Switzerland.

Kong, L.; Ma, M. Q.; Chen, G.; Xing, E. P.; Chi, Y.; Morency, L.-P.; and Zhang, K. 2023. Understanding Masked Autoencoders via Hierarchical Latent Variable Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7918–7928.

Kornblith, S.; Norouzi, M.; Lee, H.; and Hinton, G. 2019. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, 3519–3529. PMLR.

Li, W.-H.; Liu, X.; and Bilen, H. 2021. Universal Representation Learning From Multiple Domains for Few-Shot Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9526–9535.

Liu, J.; Song, L.; and Qin, Y. 2020. Prototype rectification for few-shot learning. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, 741–756. Springer.

Liu, Y.; Zhang, S.; Chen, J.; Yu, Z.; Chen, K.; and Lin, D. 2023. Improving Pixel-based MIM by Reducing Wasted Modeling Capability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5361–5372.

Liu, Y.; Zou, Y.; Li, R.; and Li, Y. 2024. Spectral Decomposition and Transformation for Cross-domain Few-shot Learning. *Neural Networks*, 179: 106536.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Mohanty, S.; Hughes, D.; and Salathé, M. 2016. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7(September). Publisher Copyright: © 2016 Mohanty, Hughes and Salathé.

Phoo, C. P.; and Hariharan, B. 2021. Self-training For Few-shot Transfer Across Extreme Task Differences. In *International Conference on Learning Representations*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *Proceedings of the International Conference on Neural Information Processing Systems*, 4080–4090.

Tseng, H.-Y.; Lee, H.-Y.; Huang, J.-B.; and Yang, M.-H. 2020. Cross-domain few-shot classification via learned feature-wise transformation. In *Proceedings of the International Conference on Learning Representations*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; and Wierstra, D. 2016. Matching networks for one shot learning. In *Proceedings of the International Conference on Neural Information Processing Systems*, 3637–3645.

Walsh, R.; Osman, I.; and Shehata, M. S. 2023. Masked Embedding Modeling With Rapid Domain Adjustment for Few-Shot Image Classification. *IEEE Transactions on Image Processing*, 32: 4907–4920.

Wang, H.; and Deng, Z.-H. 2021. Cross-Domain Few-Shot Classification via Adversarial Task Augmentation. In Zhou, Z.-H., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 1075–1081. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2022. SimMIM: A Simple Framework for Masked Image Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9653–9663.

Zhang, Z.; Chen, G.; Zou, Y.; Huang, Z.; Li, Y.; and Li, R. 2024. MICM: Rethinking Unsupervised Pretraining for Enhanced Few-shot Learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 7686–7695. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706868.

Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; and Kong, T. 2022. Image BERT Pre-training with Online Tokenizer. In *International Conference on Learning Representations*.

Zou, Y.; Liu, Y.; Hu, Y.; Li, Y.; and Li, R. 2024. Flatten Long-Range Loss Landscapes for Cross-Domain Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23575–23584.