Channel-masked Asymmetric Distribution Matching for Cross-Domain Generalized Dataset Distillation

Qi Liu[1], Chenghao Xu[1], Jiexi Yan[2*], Guangtao Lyu[1], Erkun Yang[1], Guihai Chen[1], Yanhua Yang[2*]

[1]School of Electronic Engineering, Xidian University, Xi'an, Shaanxi, China
[2]School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, China
{qiliu,chx,guangtaolyu}@stu.xidian.edu.cn,{jxyan1995,erkunyang,yanhuayang.xd}@gmail.com, ghchenwy@163.com

Abstract

Dataset distillation has achieved remarkable progress as an effective approach for data compression. However, real-world data often comes from diverse domains, leading to potential mismatches between the domains of synthesized images and those of the evaluation set. Existing methods primarily assume domain alignment between them, which limits their generalization ability in the above cross-domain scenarios. In this paper, we aim to ensure that images synthesized from known domains maintain robust performance on unseen domains and propose a novel framework called Channel-masked Asymmetric Distribution Matching (CADM). During asymmetric distribution matching, domain-sensitive channels of real data are selectively masked at different layers to extract domain-invariant features that guide synthetic data optimization. To further improve synthetic data representation, we introduce a class-focused domain-agnostic regularization to capture class-relevant knowledge while ignoring domain-specific information. Experiments show that our method produces domain-robust synthetic data and substantially improves generalization performance on unseen domains.

Introduction

Dataset distillation (DD) (Wang et al. 2018) has emerged as a promising paradigm for reducing the size of training datasets by synthesizing a compact set of informative images capable of effectively training deep neural networks. With carefully designed objectives, distilled data can achieve performance competitive with or even superior to full datasets, while significantly lowering memory and computational costs (Wang et al. 2025). These advances offer new opportunities for efficient training, rapid prototyping, and continual learning in resource-constrained settings.

However, current distillation methods predominantly assume that the domain distribution of synthetic images is consistent with that of the validation set, while real-world datasets span diverse domains with significant variations in distribution, semantics, acquisition conditions, and visual styles. For instance, object recognition datasets may encompass photos, sketches, cartoons, and paintings; medical images may originate from different scanning equipment; and autonomous driving datasets may capture various weather and lighting conditions. In such scenarios, DD needs to compress heterogeneous data from multiple domains, and simultaneously, the synthetic data will also encounter unseen domains during validation. While current DD strategies predominantly focus on in-domain scenarios. They tend to overfit to statistically dominant domains while providing insufficient generalization to out-of-domain test sets—particularly problematic under conditions of severe domain imbalance or when discriminative features exhibit strong domain-specific characteristics. When models trained on distilled images are evaluated on novel domains, their performance typically exhibits substantial degradation, as illustrated in Fig. 1(a).

Although domain generalization has made notable progress, its direct integration into dataset distillation remains challenging. Naive combinations often result in suboptimal performance. To explore this limitation, we conduct experiments using maximum mean discrepancy (MMD)-based distribution matching across several representative methods. FACT (Xu et al. 2021) augments images by mixing styles from different domains using phase and amplitude information in the fourier domain. However, such image-level perturbations yield only marginal improvements to the distilled results, suggesting limited compatibility with the distillation objective. IRM (Arjovsky et al. 2019) improves robustness by enforcing invariant predictions across domains, helping the pretrained model resist domain shifts. However, it does not directly optimize the synthetic data themselves, resulting in limited improvement. DAM (Choi et al. 2025) first focuses on multi-domain dataset distillation, mainly targeting intra-domain performance drops, but offers limited exploration of generalization under domain shifts.

Due to the aforementioned limitations, we revisit the generalization of distilled images from a novel feature-channel perspective. The central hypothesis is that the generalizability of distilled data is correlated with the robustness of specific feature channels to domain shifts. Specifically, we define activation scores to measure each channel's sensitivity to a particular domain, then quantify channel robustness in synthetic data by computing the standard deviation of these scores across different domains. As shown in Fig. 1(b), MMD-based synthetic data often contains many non-robust channels with unstable activation scores, indicating that they capture domain-specific information. When presented with

*Corresponding author.

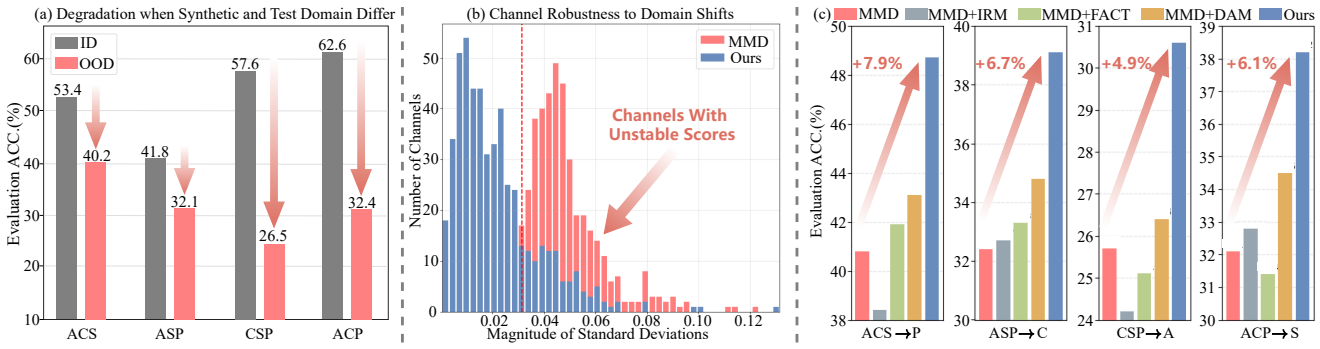

Figure 1: (a) Images synthesized on multi-domains are tested on both in-distribution and out-of-distribution data, revealing severe performance degradation on unseen domains. **ACS** represents synthetic data whose domains are sourced from **art painting**, **cartoon**, and **sketch** of PACS (Li et al. 2017). (b) Variance of channel activation scores across domains for MMD and our synthetic images. Experiments on PACS with photo as target domain, analyzing ResNet-18's last residual block representations. (c) Generalization performance comparison between our method and others. **ACS→P** denotes testing on the **photo** using synthetic data from **art painting**, **cartoon**, and **sketch**.

unseen domains, these domain-sensitive channels produce abnormal activation patterns, causing distribution shifts. We attribute this to real data providing misleading guidance during optimization due to their domain-sensitive features.

To address this challenge, we propose a novel distribution matching framework called **Channel-masked Asymmetric Distribution Matching (CADM)** for distilling images with enhanced cross-domain generalization. During pretraining, we train domain discriminators at each network layer to identify domain-sensitive channels for the distillation process. During the asymmetric distribution matching, we randomly mask highly sensitive channels in real data based on their activation scores, ensuring that real data can provide domain-invariant feature distributions to guide synthetic data optimization, while synthetic data are directly fed to subsequent layers. Furthermore, we enhance the regularization term of dataset distillation by introducing a class-focused domain-agnostic regularization constraint. It applies domain-sensitive channel masking when computing classification losses for synthetic data, while simultaneously enforcing consistency between logits obtained from the same synthetic data under different random domain-sensitive channel masking strategies. This generates more class-relevant yet domain-invariant synthetic data, improving domain robustness while preserving diversity. As illustrated in Fig. 1(c), compared to existing approaches, our method significantly improves the generalization performance of distilled images.

Our contributions are summarized as follows:

- We deeply investigate multi-domain dataset distillation and the generalization of synthetic images when the domain distribution differs between the distillation and validation phases.
- We propose CADM that uses asymmetric distribution matching and class-focused domain-agnostic regularization to enhance synthetic data robustness.
- Our method achieves state-of-the-art performance, surpassing prior approaches in producing distilled datasets with strong cross-domain generalization while maintaining competitive within-domain performance.

Related Work

Distribution Matching in Dataset Distillation

Dataset distillation was first proposed by (Wang et al. 2018; Sajedi et al. 2023) to synthesize compact training sets. Compared with bi-level optimization methods (Cazenavette et al. 2022; Shin, Shin, and Moon 2023; Xu et al. 2023), distribution matching (DM) (Zhao and Bilen 2022) balances performance and computational efficiency without nested model optimization. DM can be classified into point-wise and moment-wise matching. Moment-wise matching like DM (Zhao and Bilen 2022), IDM (Zhao et al. 2023), IID (Deng et al. 2024) minimize the maximum mean discrepancy (MMD) between synthetic and real sets, while point-wise methods designed to match features across CNN layers like DC (Zhao, Mopuri, and Bilen 2020), DSA (Zhao and Bilen 2021), DCC (Lee et al. 2022). NCFM (Wang et al. 2025) demonstrates that DM achieves superior performance with low computational overhead. However, DM may learn domain-specific distributions rather than class-discriminative patterns, particularly in multi-domain scenarios where domain-sensitive features can dominate representations and hinder generalization.

Domain Generalization

Domain generalization aims to train models that perform well on unseen domains by mitigating domain shifts without access to target domain during training. Existing methods can be broadly categorized into four groups. The image-level augmentation (Carlucci et al. 2019; Xu et al. 2021; Zhou et al. 2021; Xu et al. 2024a), enhance images diversity by altering styles or structures. Adversarial approaches (Sicilia, Zhao, and Hwang 2023; Shankar et al. 2018; Li et al. 2018) generate or adapt features to be indistinguishable across domains, thereby promoting domain invariance. Architectural strategies (Lee, Bae, and Kim 2023; Huang

et al. 2020; Yan et al. 2024b; Xu et al. 2024b) modify network structures to separate domain-invariant and domain-specific features, improving transferability. Gradient-based methods (Foret et al. 2020; Zhuang et al. 2022; Shin et al. 2024; Xu, Yan, and Deng 2025; Yan et al. 2024a) encourage gradient alignment across domains to stabilize learning and enhance generalization. While effective, these methods primarily aim to improve model robustness or directly enhance images themselves, making them less applicable to DD with limited benefits for synthetic image generalization. Our approach addresses synthetic image generalization from a distillation perspective, achieving significant improvements.

Methodology

Preliminaries

Given a real dataset $\mathcal{T} = \{(x_t^i, y_t^i)\}_{i=1}^{|\mathcal{T}|}$, where y_t^i denote the class labels, our objective is to synthesize a significantly smaller dataset $\mathcal{S} = \{(x_s^i, y_s^i)\}_{i=1}^{|\mathcal{S}|}$ that preserves critical information from $\mathcal{T}$ and enables models trained on $\mathcal{S}$ to generalize effectively on unseen target domains. Distribution matching (DM) was first introduced by (Zhao and Bilen 2022) as an alternative to traditional bi-level optimization techniques. According to NCFM (Wang et al. 2025), DM achieves strong performance with high accuracy and low memory consumption. Under identical settings, we find that replacing NCFM's complex characteristic functions and sampling networks with simple MMD loss achieves similar performance while surpassing traditional SOTA methods (Guo et al. 2023a; Shao et al. 2024a; Sun et al. 2024a).[1] We refer to this as MMD for concise presentation and adopt it as our baseline. To apply MMD, the expert model $\mathcal{M}$ is first pre-trained on the real set using a cross-entropy loss:

$$\mathcal{L}_{\text{cls}}(\mathcal{M}) = \text{CE}(\mathcal{M}(\mathcal{T}), y_t). \tag{1}$$

For distribution matching, MMD aligns moments directly in feature space like most DM-based methods (Zhao and Bilen 2022; Zhao et al. 2023; Deng et al. 2024) as:

$$\mathcal{L}_{\text{mmd}}(\mathcal{S}) = \left\| \mathbb{E}_{x_t \sim \mathcal{T}} \left[\mathcal{F}(x_t) \right] - \mathbb{E}_{x_s \sim \mathcal{S}} \left[\mathcal{F}(x_s) \right] \right\|^2, \tag{2}$$

where $\mathcal{F}$ denotes the feature extractor of $\mathcal{M}$. In addition to distribution alignment, a classification loss for the synthetic data is employed as a regularization term (Zhao et al. 2023; Yin, Xing, and Shen 2023; Wang et al. 2025). This implicitly facilitates higher-order moment alignment between real and synthetic distributions (Zhao et al. 2023) and helps ensure that synthetic features remain classically discriminative. Formally, the regularization loss is defined as:

$$\mathcal{L}_{\text{reg}}(\mathcal{S}) = \text{CE}(\mathcal{M}(\mathcal{S}), y_s). \tag{3}$$

Although MMD demonstrates excellent performance, the synthetic data it generates are severely affected by domain-specific features in the real data. When real sample features contain domain-sensitive information, the distribution matching process fails to provide effective guidance for synthesizing domain-invariant representations. Consequently, synthetic data generated through MMD struggle to achieve satisfactory generalization performance.

[1]Experiment is provided in the *supplementary materials*.

Channel-masked Asymmetric Distribution Matching

To synthesize images that achieve optimal generalization, we propose Channel-masked Asymmetric Distribution Matching (CADM), a novel approach that addresses domain-specific interference in dataset distillation. CADM introduces domain discriminators at multiple intermediate layers to identify domain-sensitive channels, then employs asymmetric distribution matching where domain-sensitive channels are selectively masked from real features while synthetic features remain unmodified and are aligned through matching. Additionally, we introduce class-focused domain-agnostic regularization that masks domain-sensitive channels of synthetic data for classification regularization, while enforcing prediction consistency under different random mask conditions. The framework is shown in Fig. 2.

Pretraining. To explicitly guide the model in removing domain-specific features during distribution matching, we introduce domain discriminators to multiple middle layers for locating domain-sensitive channels, which consists of a global average pooling (GAP) layer and a fully-connected (FC) layer. Given an input x_t^i from $\mathcal{T}$ and its domain label $\hat{y}_t^i$, we first extract the feature $\mathcal{F}_l(x_t^i) \in \mathbb{R}^{C \times H \times W}$ that is yielded by the l-th middle layer, where C is the number of channels, H and W denote the height and width dimensions, respectively. The feature map $\mathcal{F}_l(x_t^i)$ is fed to domain discriminator $\mathcal{F}_d^l$ to predict domain labels and compute discrimination loss. To avoid the negative impact of domain discriminators on the main network, we use a gradient reversal layer (GRL) (Matsuura and Harada 2020) before the domain discriminator to truncate the gradients of minimizing discrimination loss:

$$\mathcal{L}_{\text{dom}}(\mathcal{M}) = \frac{1}{L} \sum_{l=1}^{L} \text{CE}\left(\mathcal{F}_d^l(\mathcal{F}_l(x_t^i)), \hat{y}_t^i \right), \tag{4}$$

where the overall domain loss is computed as the average of domain discrimination losses across all L layers.

Asymmetric Distribution Matching. After pretraining, we obtain discriminators that can effectively distinguish image domains. During distribution matching, our objective is to encourage synthetic data to encode domain-invariant features. However, features extracted from real data often contain domain-specific information, which may hinder effective alignment. To address this, we propose asymmetric distribution matching by identifying and masking domain-sensitive channels of real data features to obtain domain-invariant distribution for guiding synthetic data.

To determine which channels encode domain-specific information, we leverages the outputs of domain discriminators. The underlying hypothesis posits that channels contributing most to accurate domain prediction are likely to encode domain-specific cues. The degree of domain specificity for each channel is quantified by computing its weighted activation with respect to the correct domain prediction, utilizing the discriminator response as an indicator of channel importance. For an real data input x_t^i and intermediate layer feature extractor $\mathcal{F}_l$, the **activation score** of the j-th channel

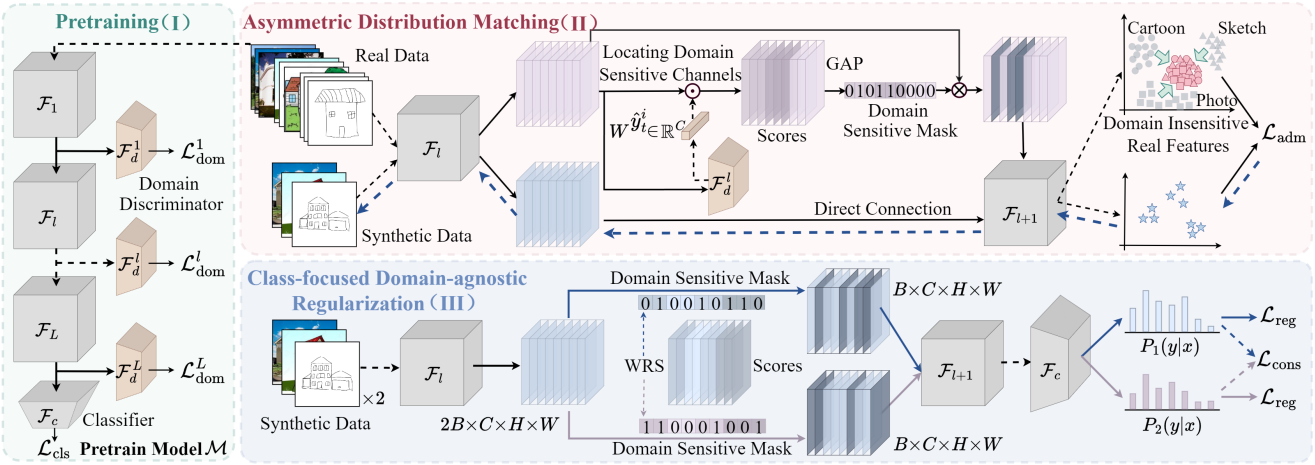

Figure 2: The overall framework of our method. Asymmetric distribution matching leverages a pretrained domain discriminator to selectively mask domain-sensitive channels of real data during the matching process, thereby extracting domain-invariant features to guide the optimization of synthetic data. Furthermore, a dual consistency regularization constraint is introduced to encourage synthetic data to learn representations that are both class-relevant and domain-robost.

in feature map $\mathcal{F}_l(x_t^i)$ is defined as:

$$s_j = W^{\hat{y}_t^i} \cdot \text{GAP}(\mathcal{F}_l(x_t^i))_j, \tag{5}$$

where $W^{\hat{y}_t^i} \in \mathbb{R}^C$ represents the FC layer weight of domain discriminator $\mathcal{F}_d$ for true domain $\hat{y}_t^i$, and C is the channel number. Higher weighted activation scores indicate greater channel contribution to domain prediction. To prevent synthetic data from overfitting to domain-specific cues, domain-sensitive channels in real data are suppressed by constraining discriminator-exploited domain information. The most domain-sensitive channels are selected and masked during matching to reduce domain-specific information in feature maps. To achieve efficient masking, the weighted random selection (WRS) algorithm (Efraimidis and Spirakis 2006) is employed. For the j-th channel with score s_j, a random number $r_j \in (0, 1)$ is generated and key value $k_j = r_j^{1/s_j}$ is computed. The mask is then set as:

$$m_j = \begin{cases} 0, & \text{if } j \in \text{TOP}(\{k_1, k_2, ..., k_C\}, M) \\ 1, & \text{otherwise} \end{cases}, \tag{6}$$

where $\text{TOP}(\{k_1, k_2, ..., k_C\}, M)$ denotes the M items with largest key values. For the current layer, we use hyper-parameter P_{active} to control the probability of performing masking in this layer, while $P_{\text{mask}} = M/C$ controls the ratio of channels masked.

Having obtained domain-insensitive real features whose distribution no longer reflects domain-specific characteristics, we perform asymmetric distribution matching between the real and the synthetic data features. We define $\widehat{\mathcal{F}}$ as the features extracted after domain-sensitive channel masking. The asymmetric distribution matching loss is formulated as:

$$\mathcal{L}_{\text{adm}}(\mathcal{S}) = \left\| \mathbb{E}_{x_t \sim \mathcal{T}}[\widehat{\mathcal{F}}(x_t)] - \mathbb{E}_{x_s \sim \mathcal{S}}[\mathcal{F}(x_s)] \right\|^2. \tag{7}$$

Now, the synthetic data is encouraged to express domain-invariant features, effectively reducing channel instability from domain shifts and promoting stable representation.

Class-focused Domain-agnostic Regularization. While the aforementioned asymmetric distribution matching promotes domain-invariant supervision for synthetic data, the regularization classification loss as Eq. 3 may still suffer from the influence of domain-specific cues, especially when such cues dominate the feature representations, thereby hindering synthetic data from learning class-focused discriminative information. To further mitigate this issue, we propose a class-focused domain-agnostic regularization strategy that enforces the synthetic data to retain class-discriminative yet domain-invariant features during training. Specifically, we extend the sensitive channel masking to the features of synthetic data when computing the classification regularization loss:

$$\mathcal{L}_{\text{reg}}(\mathcal{S}) = \text{CE}(\mathcal{F}_c(\widehat{\mathcal{F}}(x_s^i)), y_s), \tag{8}$$

where $\mathcal{F}_c$ denotes the classifier, and $\widehat{\mathcal{F}}(x_s^i)$ represents the synthetic features after masking. Through this approach, we enable synthetic data to focus on class-relevant content in classification regularization by removing domain-sensitive feature representations. To further enhance the domain robustness of synthetic data, we propose a dual consistency regularization loss. We implement the masking strategy twice with independently sampled masks for the same synthetic image input x_s^i, yielding two perturbed feature representations. These perturbations produce two potentially inconsistent prediction distributions, denoted as $\widehat{\mathcal{F}}(x_s^i)_1$ and $\widehat{\mathcal{F}}(x_s^i)_2$, respectively. This stochastic process reflects the diversity of domain-specific feature removal and can be approximated as injecting multiplicative noise (Srivastava et al. 2014; Park and Kwak 2016). We encourage the model to output consistent predictions under these two different perturbations by minimizing the symmetric KL divergence:

$$\begin{aligned} \mathcal{L}_{\text{cons}}(\mathcal{S}) = & \frac{1}{2}(\text{KL}[\sigma(\mathcal{F}_c(\widehat{\mathcal{F}}(x_s^i))_1) || \sigma(\mathcal{F}_c(\widehat{\mathcal{F}}(x_s^i))_2)] \\ & + \text{KL}[\sigma(\mathcal{F}_c(\widehat{\mathcal{F}}(x_s^i))_2) || \sigma(\mathcal{F}_c(\widehat{\mathcal{F}}(x_s^i))_1)]), \end{aligned} \tag{9}$$

Dataset	PACS								VLCS							
Domain	**ACS→P**		**ASP→C**		**CSP→A**		**ACP→S**		**SVL→C**		**SVC→L**		**SLC→V**		**VLC→S**	
IPC	1	10	1	10	1	10	1	10	1	10	1	10	1	10	1	10
Random	10.1	29.2	6.3	16.8	6.7	12.9	10.1	18.3	7.8	18.5	11.2	19.7	10.4	14.9	9.0	16.8
MMD	16.7	40.8	14.9	32.4	10.2	25.7	4.8	32.1	18.3	28.6	16.2	35.2	17.9	22.1	16.6	30.4
MMD+IRM	17.9	38.4	16.1	32.7	11.8	24.2	5.2	32.8	19.7	29.1	16.5	37.4	17.2	24.6	20.1	30.2
MMD+FACT	16.8	41.9	16.2	33.3	10.9	25.1	5.7	31.4	20.8	28.7	17.6	38.9	15.5	26.2	21.4	32.8
MMD+DAM	22.7	43.1	18.0	34.8	12.6	26.4	6.0	34.2	21.5	30.9	17.3	38.7	20.1	27.8	20.8	32.6
CADM	**25.4**	**48.7**	**21.2**	**39.1**	**14.1**	**32.6**	**15.2**	**38.2**	**21.8**	**31.3**	**20.5**	**44.3**	**25.7**	**33.7**	**21.1**	**34.8**
Full Dataset	61.1		52.2		40.8		43.2		73.3		56.7		50.4		56.6	

Office-Home								DomainNet-Sub											
ACP→R		**ACR→P**		**ARP→C**		**CRP→A**		**CIPQR→S**		**CIPQS→R**		**CIPRS→Q**		**CIQRS→P**		**CPQRS→I**		**IPQRS→C**	
1	10	1	10	1	10	1	10	1	10	1	10	1	10	1	10	1	10	1	10
5.8	6.2	4.1	4.8	4.9	5.4	4.8	5.7	3.2	4.1	3.2	3.4	3.0	4.6	3.7	6.3	2.4	3.8	3.3	5.9
6.9	11.8	5.2	8.4	5.1	8.9	4.3	9.2	3.7	5.8	3.9	3.7	4.8	4.9	3.9	8.4	2.8	4.6	4.1	9.8
7.2	12.4	5.1	8.1	6.8	9.7	4.7	11.3	4.9	6.2	3.1	4.1	3.1	4.2	4.2	8.1	3.1	5.2	4.4	10.8
7.1	11.1	4.1	9.9	6.0	9.3	3.9	10.9	4.1	6.7	4.0	4.4	4.0	4.8	4.6	8.7	3.4	5.1	5.7	11.6
8.1	13.2	6.0	9.7	**7.4**	10.6	4.0	11.6	4.1	6.6	**4.2**	4.8	4.0	4.3	4.2	9.1	3.0	6.3	5.1	11.8
10.4	**16.9**	**6.8**	**12.0**	7.3	**10.7**	**5.1**	**14.8**	**5.3**	**7.8**	4.1	**5.9**	**5.1**	**6.7**	**4.9**	**10.2**	**3.7**	**6.5**	**9.2**	**14.1**
28.8		30.3		16.7		22.0		13.0		6.7		5.4		18.8		10.4		20.9	

Table 1: The leave-one-domain-out strategy is employed to evaluate the generalization of synthetic images across four datasets. Domain abbreviations are as follows: (V)OC, (L)abelMe, (C)altech, (S)un for VLCS; (A)rt, (R)eal, (P)roduct, (C)lipart for Office-Home; and (C)lipart, (I)nfograph, (P)ainting, (Q)uickdraw, (R)eal, (S)ketch for DomainNet. IPC denotes images per class. The Full Dataset setting uses all source domains for training and evaluates on the target domain.

where σ is the softmax function. With this consistency constraint, the synthetic data is encouraged to improve the robustness of feature channels to domain shifts and extract domain-invariant features from perturbed representations. Through the synergy of these two regularization strategies, we can further obtain domain-invariant yet class-relevant feature representations that build upon the asymmetric matching foundation.

Training Scheme

The training process of our method consists of three main parts. In the pretraining phase, we jointly optimize the original model and additional domain discriminators with classification and domain discrimination losses on the real dataset to identify domain-sensitive channels across multiple intermediate layers:

$$\mathcal{L}_{\text{I}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{dom}}. \tag{10}$$

During matching, we minimize the asymmetric distribution matching loss between domain-invariant real features and synthetic features, thereby suppressing synthetic data domain-specific feature expression:

$$\mathcal{L}_{\text{II}} = \mathcal{L}_{\text{adm}}. \tag{11}$$

For regularization, we incorporate both the classification regularization loss $\mathcal{L}_{reg}$ for synthetic data which applies domain-sensitive channels masking, and the dual consistency loss $\mathcal{L}_{cons}$ to enforce class-relevant and domain-invariant representations for synthetic data:

$$\mathcal{L}_{\text{III}} = \mathcal{L}_{\text{reg}} + \lambda \mathcal{L}_{\text{cons}}. \tag{12}$$

Here, λ is a balancing hyper-parameter. To prevent excessive information removal that could hinder feature representation, we employ a layer-wise training strategy that randomly selects a single intermediate layer at each iteration to apply the channel masking operation, enabling effective domain gap reduction across the entire network hierarchy while maintaining training stability by masking domain-sensitive information from both high-level and low-level semantic features across all network layers.

Experiments

Experimental Settings

Datasets. We evaluate our method on four conventional multi-domain datasets: (1) **PACS** (Li et al. 2017) consists of images from 4 domains: Photo, Art Painting, Cartoon, and Sketch, including 7 object categories and $9,991$ images total. We adopt the official split provided by (Li et al. 2017) for training and validation. (2) **VLCS** (Torralba and Efros 2011) comprises 5 categories selected from 4 domains, VOC 2007 (Pascal), LabelMe, Caltech and Sun. We use the same setup as (Carlucci et al. 2019) and divide the dataset into training and validation sets based on $7:3$. (3) **Office-Home** (Venkateswara et al. 2017) contains around $15,500$ images of 65 categories from 4 domains: Artistic, Clipart, Product and Real-World. As in (Carlucci et al. 2019), we randomly split each domain into 90% for training and 10% for validation. (4) **DomainNet-Sub** is a subset of the large-scale DomainNet (Peng et al. 2019) dataset, consisting of 100 classes selected from the original 345 classes across 6 domains, *i.e.*, Clipart, Infograph, Painting, Quickdraw, Real, and Sketch. Following (Gulrajani and Lopez-Paz 2020), we

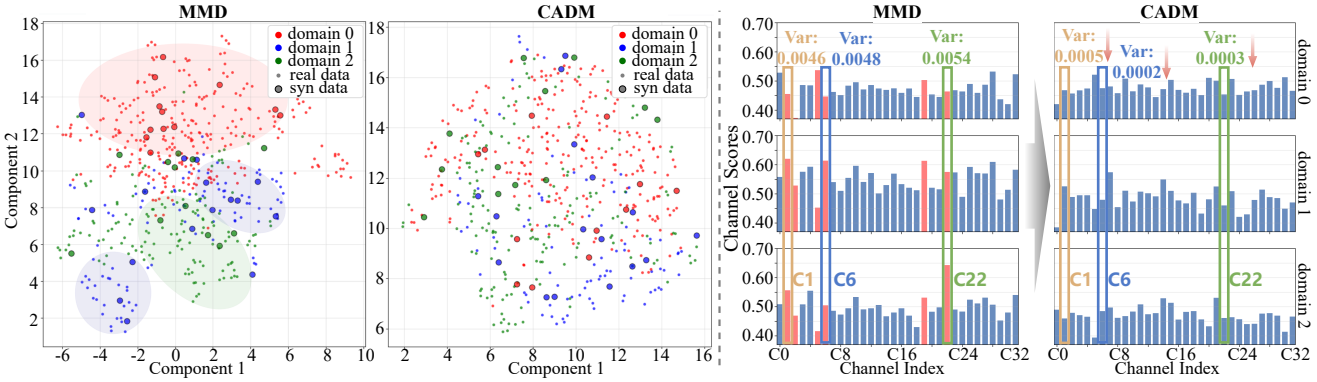

Figure 3: (a) UMAP of real vs. synthetic features (same class) under both methods, with colors indicating domains. (b) Activation scores of the first 32 channels in the last ResNet-18 block across domains. Red marks domain-sensitive channels.

split the source data into 80% training and 20% validation.

Baselines. We adopt the efficient and superior MMD-based distribution matching method as our baseline. Our experiments in the supplementary materials demonstrate that MMD, which eliminates the need for complex feature functions and sampling networks while maintaining all other settings consistent with NCFM (Wang et al. 2025), outperforms traditional state-of-the-art methods such as DATM (Guo et al. 2023b), G-VBSM (Shao et al. 2024b), and RDED (Sun et al. 2024b) on CIFAR-10 and CIFAR-100 (Krizhevsky, Hinton et al. 2009). This highlights the simplicity and effectiveness of using MMD as a baseline. For experimental fairness, we incorporate classic domain generalization methods based on data augmentation (FACT (Xu et al. 2021)) and model optimization (IRM (Arjovsky et al. 2019)) into our baseline. In addition, we extend the multi-domain dataset distillation method DAM (Choi et al. 2025) for a more comprehensive comparison.

Implementation Details. We adopt ResNet-18 (He et al. 2016a) as the backbone network, which consists of four residual blocks. Each residual block is followed by a domain discriminator implemented as a single linear layer, where the input dimension matches the number of channels, and the output dimension corresponds to the number of domains. A gradient reversal layer (GRL) (Matsuura and Harada 2020) with a weight of 0.25 is inserted before each domain discriminator. Note that GRL is only applied during the pretraining stage and is not used when optimizing synthetic data. For pretraining, we set the number of epochs to 60 for PACS and VLCS, and to 120 for Office-Home and DomainNet-Sub. We initialize synthetic data using real samples and maintain an equal number of synthetic data for each domain to reduce potential bias introduced by domain imbalance. During each matching and regularization step, we randomly select one residual block and apply masking to its feature maps. For domain-sensitive channel masking, the masking ratio P_{mask} is set to 0.33 and P_{active} is set to 0.8. λ is set to 100. Following NCFM (Wang et al. 2025), we adopt differential augmentation (Zhao and Bilen 2021; Wang et al. 2022) and employ multi-formation parameterization with a scale factor of $\rho = 2$ for image inputs, as described in (Kim et al. 2022; Zhao et al. 2023). All experiments are implemented using PyTorch (Paszke et al. 2017) and conducted on an NVIDIA A6000 GPU.

Experimental Results

Quantitative Analysis. Tab. 1 demonstrate the generalization capability of synthetic data generated by our method across four multi-domain datasets on unseen domains. Note that we employ a leave-one-out strategy for each synthesis, reserving one domain for validation. For fair comparison, we extend the MMD baseline with classic domain generalization methods IRM and FACT, as well as the multi-domain distillation method DAM. For IRM, we employ it during the pre-training stage. However, its improvement is limited as it only enhances the pre-trained model's capability without further enhancement during the distillation process. For FACT, we apply the fourier-based data augmentation strategy during both pre-training and distillation stages on real and synthetic data. Its limitation lies in merely augmenting data without effectively fusing knowledge from different domains during distillation. In some cases, the augmented synthetic data even produce adverse effects. DAM can distill multi-domain datasets but primarily targets in-domain performance enhancement, showing limited generalization capability for synthetic data. In contrast, our method exhibits substantial performance across multiple datasets, demonstrating its effectiveness in distilling domain-invariant images while preserving feature diversity. This enables the synthetic data to capture domain-robust representations, thereby achieving strong performance on unseen domains.

Qualitative Analysis. To analyze the domain invariance of synthetic data, we conduct UMAP (McInnes, Healy, and Melville 2018) visualization analysis on both synthetic and real data features obtained from MMD and CADM methods, as shown in Fig.3(a). For the same class, MMD's real data exhibit clear discrimination across different domains, and correspondingly, the synthetic data distributions also demonstrate strong domain correlation. In contrast, due to the removal of domain-sensitive channels, CADM's real data features exhibit better domain invariance, and the synthetic data similarly present more uniform domain-invariant

$\mathcal{L}_{adm}$	$\mathcal{L}_{reg}$	$\mathcal{L}_{cons}$	IPC	
			1	10
-	-	-	16.4	39.9
✓	-	-	23.6	45.9
-	✓	-	21.4	40.7
-	✓	✓	23.1	43.3
✓	✓	✓	**25.1**	**48.4**

Table 2: Ablation experiments on proposed loss, which are conducted under the **ACS→P** setting on PACS.

IPC	Method	ConvNet	AlexNet	VGG	ResNet
1	MMD	22.5	18.2	21.3	23.1
	CADM	**24.8**	**22.7**	**23.2**	**24.3**
10	MMD	39.3	38.8	36.1	41.2
	CADM	**45.7**	**44.4**	**43.8**	**45.9**

Table 3: Cross-architecture generalization under **ACS→P**. Images are distilled by ResNet-18 and evaluated on others. The test domain differs from all synthetic domains.

distributions. This indicates that our method enables real data to provide better guidance for distribution matching, thereby endowing synthetic data distributions with enhanced domain robustness. More visualization results of synthetic data are provided in the *supplementary materials*.

Ablation Study and Analysis

Analysis of the Effectiveness of Different Losses. Here, we analyze the effectiveness of our main proposed asymmetric distribution matching loss $\mathcal{L}_{\text{adm}}$, regularization classification loss using sensitive channel mask $\mathcal{L}_{\text{reg}}$, and dual consistency loss $\mathcal{L}_{\text{cons}}$. As shown in Tab. 2, different loss functions contribute varying degrees of performance improvement. The asymmetric distribution matching loss $\mathcal{L}_{\text{adm}}$ yields the most significant improvement of 6.0%, indicating that masking domain-sensitive channels effectively corrects the distribution of real data to better guide synthetic data. Even when applied alone, this also provides substantial performance gains. The classification and dual-consistency losses add 3.4% gain, showing that the it helps synthetic data learn class-discriminative and domain-robust features.

Activation Scores of Synthetic Data Across Different Domains. We analyze the activation scores of the last layer channels for synthetic data generated by MMD and our CADM across different domains under the same synthetic data initialization. The activation score represents the sensitivity of a channel to a specific domain. As shown in Fig. 3(b), the red channels in MMD represent channels of synthetic data with large variances across different domains, which are the unstable channels shown in Fig. 1(b) and are typically most affected by domains. Meanwhile, the mean activation scores across different domains also exhibit significant differences. In contrast, our CADM significantly reduces the number of unstable channels, and compared to MMD, the variance of activation scores for sensitive channels is substantially decreased, indicating that we successfully suppress the expression of domain-sensitive channels in synthetic data. Furthermore, the mean activation scores across different domains remain largely consistent, indicating that the channel activations of synthetic data are stable across domains and exhibit domain-invariant characteristics.

Dataset	PACS		VLCS		Office-Home		DomainNet-Sub	
IPC	1	10	1	10	1	10	1	10
MMD	40.3	48.2	36.2	41.4	15.4	24.9	12.4	24.9
MMD+IRM	41.7	48.2	38.6	44.3	18.1	25.3	11.1	27.3
MMD+FACT	43.3	47.5	40.1	46.7	19.7	27.1	13.7	27.1
MMD+DAM	46.9	52.2	42.7	46.9	19.4	30.5	15.4	29.5
CADM	**49.2**	**58.4**	**43.1**	**50.4**	**24.8**	**33.5**	**17.3**	**30.6**
Full Dataset	71.3		64.8		49.2		48.3	

Table 4: The results of test models trained on distilled synthetic images under the in-domain setting, where all domains are used as both source and target domains.

In-domain Evaluation Stability of Synthetic Data. To confirm that our synthetic data improve both out-of-domain and in-domain performance, we conduct additional experiments. As shown in Tab. 4, we distill images using all domains in each dataset and evaluate on all corresponding domains. CADM consistently outperforms the MMD+DAM baseline, showing better multi-domain distillation. This verifies that our method effectively fuses cross-domain information, producing class-relevant, domain-invariant representations that enhance both in-domain stability and out-of-domain generalization.

Cross-Architecture Generalization. We assess cross-architecture generalization on ConvNet-3 (Gidaris and Komodakis 2018), AlexNet (Krizhevsky, Hinton et al. 2009; Yan et al. 2021), VGG-11 (Simonyan and Zisserman 2014; Lyu et al. 2025), and ResNet-50 (He et al. 2016b). Synthetic data are distilled with ResNet-18 and tested on each architecture. Tab. 3 reports results on PACS with 1 and 10 IPC under the **ACS→P** setting. In both cases, CADM consistently outperforms MMD across all architectures, demonstrating that our synthetic data maintains effective generalization performance across different domains when evaluated with various network.

Conclusion

To address the generalization problem of distilled images across different domains, we propose a channel-masked asymmetric distribution matching framework. We leverage pre-trained domain discriminators to identify domain-highly-correlated channels at different layers in real data, then implement a channel masking strategy to suppress the expression of these channels, thereby providing correct guidance for synthetic data. During regularization, We apply class-focused domain-agnostic regularization by designing a classification loss with domain-sensitive channel masking and a dual consistency regularization loss to enable synthetic data to acquire class-relevant but domain-agnostic features. Extensive experimental results demonstrate that synthetic images generated by our method achieve superior performance on test sets from unseen domains.

Acknowledgments

Our work is supported in part by the National Key R&D Program of China (No. 2023YFC3305600), the Joint Fund of Ministry of Education of China (8091B022149, 8091B02072404), and National Natural Science Foundation of China (62132016, 62571412, 62302372).

References

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Carlucci, F. M.; D'Innocente, A.; Bucci, S.; Caputo, B.; and Tommasi, T. 2019. Domain generalization by solving jigsaw puzzles. In *CVPR*.

Cazenavette, G.; Wang, T.; Torralba, A.; Efros, A. A.; and Zhu, J.-Y. 2022. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4750–4759.

Choi, J.; Han, G.; Lee, D.-J.; Baek, S.; and Kim, J. 2025. DAM: Domain-Aware Module for Multi-Domain Dataset Condensation. *arXiv preprint arXiv:2505.22387*.

Deng, W.; Li, W.; Ding, T.; Wang, L.; Zhang, H.; Huang, K.; Huo, J.; and Gao, Y. 2024. Exploiting Inter-sample and Inter-feature Relations in Dataset Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17057–17066.

Efraimidis, P. S.; and Spirakis, P. G. 2006. Weighted random sampling with a reservoir. *Information processing letters*.

Foret, P.; Kleiner, A.; Mobahi, H.; and Neyshabur, B. 2020. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*.

Gidaris, S.; and Komodakis, N. 2018. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4367–4375.

Gulrajani, I.; and Lopez-Paz, D. 2020. In Search of Lost Domain Generalization. In *ICLR*.

Guo, Z.; Wang, K.; Cazenavette, G.; Li, H.; Zhang, K.; and You, Y. 2023a. Towards lossless dataset distillation via difficulty-aligned trajectory matching. *arXiv preprint arXiv:2310.05773*.

Guo, Z.; Wang, K.; Cazenavette, G.; Li, H.; Zhang, K.; and You, Y. 2023b. Towards lossless dataset distillation via difficulty-aligned trajectory matching. *arXiv preprint arXiv:2310.05773*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, Z.; Wang, H.; Xing, E. P.; and Huang, D. 2020. Self-challenging improves cross-domain generalization. In *ECCV*.

Kim, J.-H.; Kim, J.; Oh, S. J.; Yun, S.; Song, H.; Jeong, J.; Ha, J.-W.; and Song, H. O. 2022. Dataset condensation via efficient synthetic-data parameterization. In *International Conference on Machine Learning*, 11102–11118. PMLR.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lee, S.; Bae, J.; and Kim, H. Y. 2023. Decompose, Adjust, Compose: Effective Normalization by Playing with Frequency for Domain Generalization. In *CVPR*.

Lee, S.; Chun, S.; Jung, S.; Yun, S.; and Yoon, S. 2022. Dataset condensation with contrastive signals. In *International Conference on Machine Learning*, 12352–12364. PMLR.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, 5542–5550.

Li, H.; Pan, S. J.; Wang, S.; and Kot, A. C. 2018. Domain generalization with adversarial feature learning. In *CVPR*.

Lyu, G.; Xu, C.; Yan, J.; Yang, M.; and Deng, C. 2025. Towards Unified Human Motion-Language Understanding via Sparse Interpretable Characterization. In *ICLR*, 1–25.

Matsuura, T.; and Harada, T. 2020. Domain generalization using a mixture of multiple latent domains. In *AAAI*.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Park, S.; and Kwak, N. 2016. Analysis on the dropout effect in convolutional neural networks. In *ACCV*.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.(2017).

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *ICCV*.

Sajedi, A.; Khaki, S.; Amjadian, E.; Liu, L. Z.; Lawryshyn, Y. A.; and Plataniotis, K. N. 2023. Datadam: Efficient dataset distillation with attention matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17097–17107.

Shankar, S.; Piratla, V.; Chakrabarti, S.; Chaudhuri, S.; Jyothi, P.; and Sarawagi, S. 2018. Generalizing Across Domains via Cross-Gradient Training. In *ICLR*.

Shao, S.; Yin, Z.; Zhou, M.; Zhang, X.; and Shen, Z. 2024a. Generalized large-scale data condensation via various backbone and statistical matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16709–16718.

Shao, S.; Yin, Z.; Zhou, M.; Zhang, X.; and Shen, Z. 2024b. Generalized large-scale data condensation via various backbone and statistical matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16709–16718.

Shin, D.; Shin, S.; and Moon, I.-C. 2023. Frequency domain-based dataset distillation. *Advances in Neural Information Processing Systems*, 36: 70033–70044.

Shin, S.; Bae, H.; Na, B.; Kim, Y.-Y.; and Moon, I.-C. 2024. Unknown Domain Inconsistency Minimization for Domain Generalization. *arXiv preprint arXiv:2403.07329*.

Sicilia, A.; Zhao, X.; and Hwang, S. J. 2023. Domain adversarial neural networks for domain generalization: When it works and how to improve. *Machine Learning*, 112(7): 2685–2721.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*.

Sun, P.; Shi, B.; Yu, D.; and Lin, T. 2024a. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9390–9399.

Sun, P.; Shi, B.; Yu, D.; and Lin, T. 2024b. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9390–9399.

Torralba, A.; and Efros, A. A. 2011. Unbiased look at dataset bias. In *CVPR*.

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep hashing network for unsupervised domain adaptation. In *CVPR*.

Wang, K.; Zhao, B.; Peng, X.; Zhu, Z.; Yang, S.; Wang, S.; Huang, G.; Bilen, H.; Wang, X.; and You, Y. 2022. Cafe: Learning to condense dataset by aligning features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12196–12205.

Wang, S.; Yang, Y.; Liu, Z.; Sun, C.; Hu, X.; He, C.; and Zhang, L. 2025. Dataset distillation with neural characteristic function: A minmax perspective. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 25570–25580.

Wang, T.; Zhu, J.-Y.; Torralba, A.; and Efros, A. A. 2018. Dataset distillation. *arXiv preprint arXiv:1811.10959*.

Xu, C.; Guangtao, L.; Jiexi, Y.; Muli, Y.; and Deng, C. 2024a. LLM Knows Body Language, Too: Translating Speech Voices into Human Gestures. In *ACL*, 14734–14751.

Xu, C.; Yan, J.; and Deng, C. 2025. Keep and Extent: Unified Knowledge Embedding for Few-Shot Image Generation. *IEEE Transactions on Image Processing*.

Xu, C.; Yan, J.; Yang, M.; and Deng, C. 2024b. Rethinking noise sampling in class-imbalanced diffusion models. *IEEE Transactions on Image Processing*.

Xu, C.; Yan, J.; Yang, Y.; and Deng, C. 2023. Implicit Compositional Generative Network for Length-Variable Co-Speech Gesture Synthesis. *IEEE TMM*, 6325–6335.

Xu, Q.; Zhang, R.; Zhang, Y.; Wang, Y.; and Tian, Q. 2021. A Fourier-based Framework for Domain Generalization. In *CVPR*.

Yan, J.; Deng, C.; Huang, H.; and Liu, W. 2024a. Causality-invariant interactive mining for cross-modal similarity learning. *IEEE TPAMI*, 1–15.

Yan, J.; Luo, L.; Deng, C.; and Huang, H. 2021. Unsupervised hyperbolic metric learning. In *CVPR*, 12465–12474.

Yan, J.; Yin, Z.; Xu, C.; Deng, C.; and Huang, H. 2024b. Retrieval across any domains via large-scale pre-trained model. In *Forty-first International Conference on Machine Learning*.

Yin, Z.; Xing, E.; and Shen, Z. 2023. Squeeze, Recover and Relabel: Dataset Condensation at ImageNet Scale From A New Perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhao, B.; and Bilen, H. 2021. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning*, 12674–12685. PMLR.

Zhao, B.; and Bilen, H. 2022. Dataset Condensation with Distribution Matching. arXiv:2110.04181.

Zhao, B.; Mopuri, K. R.; and Bilen, H. 2020. Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929*.

Zhao, G.; Li, G.; Qin, Y.; and Yu, Y. 2023. Improved distribution matching for dataset condensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7856–7865.

Zhou, K.; Yang, Y.; Qiao, Y.; and Xiang, T. 2021. Domain Generalization with MixStyle. In *ICLR*.

Zhuang, J.; Gong, B.; Yuan, L.; Cui, Y.; Adam, H.; Dvornek, N.; Tatikonda, S.; Duncan, J.; and Liu, T. 2022. Surrogate gap minimization improves sharpness-aware training. *arXiv preprint arXiv:2203.08065*.