SVasP: Self-Versatility Adversarial Style Perturbation for Cross-Domain Few-Shot Learning

Wenqian Li[1,2], **Pengfei Fang**[1,2*], **Hui Xue**[1,2*]

[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
{wenqianli.li, fangpengfei, hxue}@seu.edu.cn

Abstract

Cross-Domain Few-Shot Learning (CD-FSL) aims to transfer knowledge from seen source domains to unseen target domains, which is crucial for evaluating the generalization and robustness of models. Recent studies focus on utilizing visual styles to bridge the domain gap between different domains. However, the serious dilemma of gradient instability and local optimization problem occurs in those style-based CD-FSL methods. This paper addresses these issues and proposes a novel crop-global style perturbation method, called **S**elf-**V**ersatility **A**dversarial **S**tyle **P**erturbation (**SVasP**), which enhances the gradient stability and escapes from poor sharp minima jointly. Specifically, SVasP simulates more diverse potential target domain adversarial styles via diversifying input patterns and aggregating localized crop style gradients, to serve as global style perturbation stabilizers within one image, a concept we refer to as self-versatility. Then a novel objective function is proposed to maximize visual discrepancy while maintaining semantic consistency between global, crop, and adversarial features. Having the stabilized global style perturbation in the training phase, one can obtain a flattened minima in the loss landscape, boosting the transferability of the model to the target domains. Extensive experiments on multiple benchmark datasets demonstrate that our method significantly outperforms existing state-of-the-art methods.

Code — https://github.com/liwenqianSEU/SVasP

Introduction

Deep learning models have achieved significant advancements in visual recognition when trained with abundant labeled samples. However, in many real-world applications, such as rare disease diagnosis, large training datasets with reliable annotations are not always feasible. To address this limitation, Few-Shot Learning (FSL) methods have been developed to enable models to generalize to novel classes with only a few samples per class (Triantafillou et al. 2020; Feng et al. 2024b). In addition to the challenge of limited data, there is often a domain gap between the source domains and target domains in practical scenarios, which presents a critical challenge. Consequently, Cross-Domain Few-Shot Learning (CD-FSL) methods have been explored to transfer domain-agnostic knowledge from multiple well-annotated source domains to target domains with limited labeled data (Tseng et al. 2020; Feng, Wang, and Geng 2024). Among the various CD-FSL settings, Single Source CD-FSL addresses domain shifts more realistically by restricting the model to access only one source domain during training.

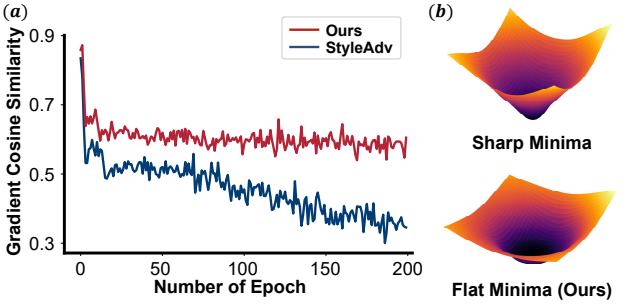

Figure 1: SVasP stabilizes the gradients and escapes from poor sharp minima. (a) demonstrates the gradient cosine similarity between epochs for displaying ground representations of gradient stability, and the larger the cosine similarity, the more stable gradient update direction. (b) demonstrates the proposed approach ensures that the model converges to a flat minima and is robust to domain shifts.

Recent studies have explored perturbing the styles of images to facilitate models acquiring more domain-agnostic knowledge from a single source domain (Kim and Han 2024; Zhong et al. 2022). By recognizing image styles (*e.g.*, mean and standard deviation) as key domain characteristics (Zhou et al. 2020), these studies aim to enhance model generalization and mitigate domain shifts by altering these domain-specific attributes (Feng et al. 2023; Xie et al. 2024). Although style-based methods have demonstrated effectiveness in Cross-Domain Few-Shot Learning (CD-FSL), they remain suboptimal due to the inherent differences between domains and the varied optimization paths for adversarial perturbations. As a result, models may overfit to noise or specific samples, becoming overly dependent on the source domain and thereby limiting their generalization capabilities.

Recently, StyleAdv (Fu et al. 2023) has addressed domain shifts by augmenting the original styles with signed gra-

*Corresponding author.

dients. Although effective in CD-FSL tasks, StyleAdv exhibits significant gradient instability. As illustrated in Figure 1 (a), we measure the gradient cosine similarity between the forward and backward gradients to assess gradient stability. However, we observe a continuous decline and severe oscillations in gradient cosine similarity, indicating that stable gradient optimization is unattainable. This instability is attributed to the absence of target domains and inadequate collection of source domain gradients, which can misdirect adversarial style attacks (Wang and He 2021). Moreover, StyleAdv's strategy of employing minimal perturbations for adversarial training tends to make the model overly sensitive to such perturbations, thereby undermining its robustness.

To address these challenges, we leverage diverse inputs from the source domain to enhance style diversity and propose a novel framework called Self-Versatility Adversarial Style Perturbation (**SVasP**). We argue that localized crop style gradients play a crucial role in model performance. The core idea of SVasP is to enhance the transferability of source domain knowledge by integrating localized crop style gradients with global style optimization. Contrary to StyleAdv, our SVasP improves the stability of model gradient in the optimization phase, shown in Figure 1 (a). During training, the issue of gradient oscillation is effectively mitigated, allowing the model to escape sharp minima and achieve smoother, flatter minima, which are more conducive to improving the model's generalization, as illustrated in Figure 1 (b).

Specifically, our method employs a structure of inner and outer iterations. In each outer iteration, sections of the benign image are randomly cropped and resized for use in the subsequent inner iterations. During each inner iteration, we iteratively generate and integrate all crop style gradients, applying them to target the global style of the benign image. The central concept of our approach is to stabilize the gradients by incorporating as much relevant gradient information from the source domain as possible. To the best of our knowledge, this is the inaugural study exploring the impact of localized style gradients on model generalization.

The main contribution of our paper is three-fold:

- We propose a new framework called SVasP that incorporates crop style gradients with the global style gradients within a image itself, which is called self-versatility, to efficiently stabilize gradients for adversarial style attack and escape from the sharp minima.
- We design a novel objective function, named Discrepancy & Consistency Optimization (DCO) to maximize visual discrepancy between seen and unseen domains while maintaining semantic consistency.
- We conduct extensive experiments on multiple benchmark datasets and validate the effectiveness of our modules. The quantitative results show that our proposed SVasP significantly improves the model's generalizability over other state-of-the-art(SOTA) methods.

Related Work

Cross-Domain Few-Shot Learning. Cross-Domain Few-Shot Learning (CD-FSL) aims to train a model on source domains that can effectively generalize to target domains, first introduced in (Chen et al. 2018a). Key benchmarks include BSCD-FSL (Guo et al. 2020), *mini*-CUB (Tseng et al. 2020), and Meta-Dataset (Triantafillou et al. 2020).

CD-FSL methods can be categorized based on access to target domain data: Single Source CD-FSL (Zou et al. 2024; Hu and Ma 2022), unlabeled target-domain CD-FSL (Islam et al. 2021; ZHENG et al. 2023), and labeled target-domain CD-FSL (Fu, Fu, and Jiang 2021). This paper focuses on the most realistic and challenging setting, Single Source CD-FSL, where only a source domain dataset is accessible.

Input Diversity for Domain Shift. To address domain shift, many methods enhance input diversity. In domain generalization, MiRe (Chen et al. 2022) mixes images from different domains, and CreTok (Feng et al. 2024a) combines tokens for creative generation. In object detection, DoubleAUG (Qi et al. 2024) exchanges RGB channels, and RECODE (Li et al. 2024) decomposes visual features into subject, object, and spatial features. In CD-FSL, LDP-net (Zhou et al. 2023) extracts local features, TGDM (Zhuo et al. 2022) and meta-FDMixup (Fu, Fu, and Jiang 2021) mix source and auxiliary data, and ConFeSS (Das, Yun, and Porikli 2021) use different augmentation methods. These augmentation methods generate diverse input patterns and more generic features for transfer. However, none of these works consider the gradient instability problem, which is a critical issue in Single Source CD-FSL.

Gradient-based Optimization. Various gradient-based optimization methods improve model robustness and generalization. GradNorm (Chen et al. 2018b) and GAM (Zhang et al. 2023) explore gradient normalization techniques. CGDM (Du et al. 2021) minimizes the discrepancy between gradients from source and target samples. Fishr (Rame, Dancette, and Cord 2022) aligns domain-level loss landscapes by leveraging gradient covariances, and PCGrad (Yu et al. 2020) addresses conflicting gradients in multi-task learning. However, these methods often overlook diverse patterns, such as crop image style gradients, which limits their effectiveness in addressing model overfitting.

Methodology

This section introduces the proposed novel framework **SVasP**, designed for CD-FSL. An overview of our method is depicted in Figure 2.

Problem Formulation

We focus on the Single Source CD-FSL setting where only a source dataset $\mathcal{D}^s$ can be accessed while the target dataset $\mathcal{D}^t$ is forbidden. Notably, for CD-FSL, $C(\mathcal{D}^s) \cap C(\mathcal{D}^t) = \emptyset$, $P(\mathcal{D}^s) \neq P(\mathcal{D}^t)$, where $C(\cdot)$ and $P(\cdot)$ denote the categories and distributions of the source and target dataset, respectively. Moreover, episode training is used in this work. Specifically, to simulate the N-way K-shot problem, N classes are selected and K samples per class are chosen to form the support set $\mathcal{S} = \{\boldsymbol{x}_i^s, y_i^s\}_{i=1}^{n_s}$, where $n_s = NK$. And the same N classes with another M images are used to construct the query set $\mathcal{Q} = \{\boldsymbol{x}_i^q\}_{i=1}^{n_q}$, where $n_q = NM$. Therefore, an episode $\mathcal{T} = (\mathcal{S}, \mathcal{Q})$ is constituded, comprising of a support set $\mathcal{S}$ and a query set $\mathcal{Q}$, and $|\mathcal{T}| =$

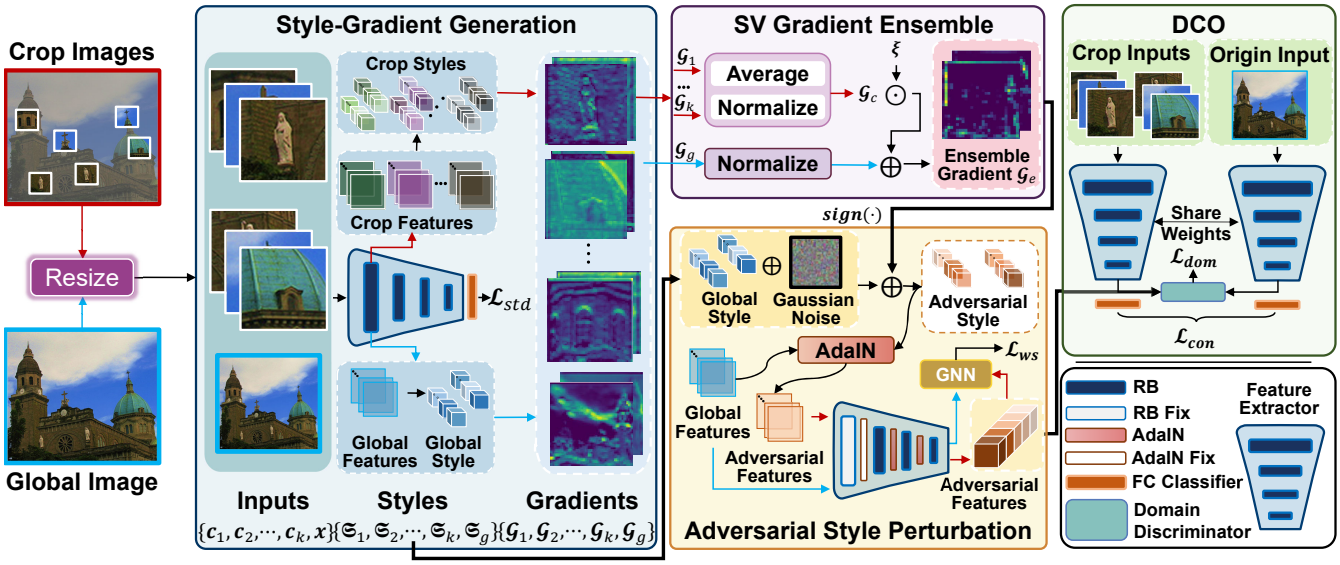

Figure 2: Overview of our proposed methods SVasP. "RB" is an abbreviation for ResNet Block. Random cropping the benign image and generates several crop images. Then, four main modules are performed: a) Generate the gradients of both crop and global styles (illustration with B_1); b) Integrate localized crop style gradients into the global style gradients; c) Perform adversarial style perturbation based on AdaIN method; d) DCO: Maximize domain visual discrepancy and global-crop consistency.

$N(K+M)$. The goal is to classify the images of the query set by training a feature extractor and a classification head on the support set.

SVasP

Overview. The proposed model contains a CNN/ViT backbone E, a domain discriminator f_{dom}, a global FC classifier f_g and a FSL relation classifier f_{re} with learnable parameter θ_E, θ_{dom}, θ_g and θ_{re}, respectively.

The network consists of four components: Style-Gradient Generation module to produce global and crop style gradients, Self-Versatility (SV) Gradient Ensemble module to integrate the localized crop style gradients as the global perturbation stabilizers, Adversarial Style Perturbation module to simulate diverse unseen styles, and Discrepancy & Consistency Optimization (DCO) to maximize the discrepancy between seen and unseen domains and maintains global-crop semantic consistency.

Without accessing auxiliary data, SVasP moderates the gradient instability and achieves a flatter minima, robustly improving the model's generalizability. Further details are provided in the following sections.

Style-Gradient Generation. In this paper, the styles of global and crop features are modeled as Gaussian distributions (Li et al. 2022b) and learnable parameters which will be updated by adversarial training. Specifically, for feature maps $\boldsymbol{F} \in \mathbb{R}^{B\times C\times H\times W}$, where B, C, H and W denote the batch size, channel, height, width of the feature maps $\boldsymbol{F}$, the specific formula for calculating the style $\mathfrak{S} = \{\boldsymbol{\mu}, \boldsymbol{\sigma}\}$ is:

$$\boldsymbol{\mu} = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\boldsymbol{F}_{B,C,h,w}, \tag{1}$$

$$\boldsymbol{\sigma} = \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(\boldsymbol{F}_{B,C,h,w}-\boldsymbol{\mu})^2 + \epsilon}, \tag{2}$$

where ϵ is a small value to avoid division by zero.

Unlike directly perturbing the global style, we consider incorporating crop style gradients to stabilize the global style gradients. For each benign image and label pair $(\boldsymbol{x}, y)$, we randomly crop k local images by the scale parameter $s = \{s_l, s_h\}$, where s_l and s_h denote the lower and upper bound for the area of the random cropped images respectively and get the input set $\mathbb{I} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_k, \boldsymbol{x}\}$. Instead of generating the adversarial style of all blocks' features at once, we use an iterative approach. Concretely, the embedding module E has four blocks B_1, B_2, B_3, B_4, and style transformation only performs on the first three blocks, as the shallow blocks produce more migratory features. For each block B_j of the backbone E, we obtain the crop and global feature maps $\mathbb{F}^j = \{\boldsymbol{F}_1^j, \boldsymbol{F}_2^j, \cdots, \boldsymbol{F}_k^j, \boldsymbol{F}_g^j\}$. For each $\boldsymbol{F}^j \in \mathbb{F}^j$, $\boldsymbol{F}^j \in \mathbb{R}^{B\times C\times H\times W}$, $\boldsymbol{F}^j$ is accumulated from block 1 to block $j-1$:

$$\boldsymbol{F}^j = \mathfrak{T}_j(\mathfrak{T}_{j-1}(\cdots(\mathfrak{T}_1(\boldsymbol{I}, \mathfrak{S}_{adv}^1), \cdots), \mathfrak{S}_{adv}^{j-1}), \mathfrak{S}_{adv}^j) \tag{3}$$

where transferring features between block $j-1$ and block j is formulated as:

$$\mathfrak{T}_j(\boldsymbol{F}^j, \mathfrak{S}_{adv}^j) = \frac{B_j(\boldsymbol{F}^{j-1}) - \boldsymbol{\mu}_{F^j}}{\boldsymbol{\sigma}_{F^j}} * \boldsymbol{\sigma}_{adv}^j + \boldsymbol{\mu}_{adv}^j \tag{4}$$

and the style $\mathfrak{S}_{F^j} = \{\boldsymbol{\mu}_{F^j}, \boldsymbol{\sigma}_{F^j}\}$ of $\boldsymbol{F}^j$ is caculated by Eq. (1) and (2). Thus, we can get the styles of the feature maps of B_j to form the style set $\mathbb{S} = \{\mathfrak{S}_1^j, \mathfrak{S}_2^j, \cdots, \mathfrak{S}_k^j, \mathfrak{S}_g^j\}$. Then, we continue to pass $\boldsymbol{F}^j$ to the remainder

of the backbone and the global FC classifier without performing any other operations and get the final prediction $\boldsymbol{p} = f_g(B_4(\cdots(B_{j+1}(\boldsymbol{F}^j)));\theta_g), \boldsymbol{p} \in \mathbb{R}^{B\times N_c}$, where N_c denotes the total number of classes. Thus the total prediction set is $\mathbb{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_k, \boldsymbol{p}_g\}$. Therefore the classification loss can be written as:

$$\mathcal{L}_{cls} = \mathcal{L}_{CE}(\boldsymbol{p}_g, y) + \sum\nolimits_{i=1}^{k} \mathcal{L}_{CE}(\boldsymbol{p}_i, y) \tag{5}$$

where $\mathcal{L}_{CE}(\cdot,\cdot)$ denotes the cross-entropy loss.

The sequel will compute the adversarial style of block B_j, we omit the subscript j for readability and calculate the gradients of the mean $\boldsymbol{\mu}$ and the std $\boldsymbol{\sigma}$ by loss back propagation:

$$\begin{aligned}\mathbb{G}^{\mu} &= \{\boldsymbol{\mathcal{G}}_1^{\mu}, \boldsymbol{\mathcal{G}}_2^{\mu}, \cdots, \boldsymbol{\mathcal{G}}_k^{\mu}, \boldsymbol{\mathcal{G}}_g^{\mu}\} \\ &= \{\nabla_{\boldsymbol{\mu}_1}\mathcal{L}_{cls}, \nabla_{\boldsymbol{\mu}_2}\mathcal{L}_{cls}, \cdots, \nabla_{\boldsymbol{\mu}_k}\mathcal{L}_{cls}, \nabla_{\boldsymbol{\mu}_g}\mathcal{L}_{cls}\}\end{aligned} \tag{6}$$

$$\begin{aligned}\mathbb{G}^{\sigma} &= \{\boldsymbol{\mathcal{G}}_1^{\sigma}, \boldsymbol{\mathcal{G}}_2^{\sigma}, \cdots, \boldsymbol{\mathcal{G}}_k^{\sigma}, \boldsymbol{\mathcal{G}}_g^{\sigma}\} \\ &= \{\nabla_{\boldsymbol{\sigma}_1}\mathcal{L}_{cls}, \nabla_{\boldsymbol{\sigma}_2}\mathcal{L}_{cls}, \cdots, \nabla_{\boldsymbol{\sigma}_k}\mathcal{L}_{cls}, \nabla_{\boldsymbol{\sigma}_g}\mathcal{L}_{cls}\}\end{aligned} \tag{7}$$

Other blocks' style gradients can be generated likewise.

SV Gradient Ensemble. Self-Versatility (SV) Gradient Ensemble module serves as the core part of our work, dedicated to bootstrapping global style gradients by integrating localized crop style gradients. We first average and normalize the style gradients of the crops to get the aggregate crop style gradients $\mathbb{G}^c = \{\boldsymbol{\mathcal{G}}_c^{\mu}, \boldsymbol{\mathcal{G}}_c^{\sigma}\}$, where:

$$\boldsymbol{\mathcal{G}}_c^{\mu} = Norm(\frac{1}{k}\sum(\boldsymbol{\mathcal{G}}_1^{\mu} + \boldsymbol{\mathcal{G}}_2^{\mu} + \cdots + \boldsymbol{\mathcal{G}}_k^{\mu})) \tag{8}$$

$$\boldsymbol{\mathcal{G}}_c^{\sigma} = Norm(\frac{1}{k}\sum(\boldsymbol{\mathcal{G}}_1^{\sigma} + \boldsymbol{\mathcal{G}}_2^{\sigma} + \cdots + \boldsymbol{\mathcal{G}}_k^{\sigma})) \tag{9}$$

Subsequently, a decay factor ξ is introduced to finally get the ensemble style gradients $\mathbb{G}^e = \{\boldsymbol{\mathcal{G}}_e^{\mu}, \boldsymbol{\mathcal{G}}_e^{\sigma}\}$, where:

$$\boldsymbol{\mathcal{G}}_e^{\mu} = Norm(\boldsymbol{\mathcal{G}}_g^{\mu}) + \xi \odot \boldsymbol{\mathcal{G}}_c^{\mu} \tag{10}$$

$$\boldsymbol{\mathcal{G}}_e^{\sigma} = Norm(\boldsymbol{\mathcal{G}}_g^{\sigma}) + \xi \odot \boldsymbol{\mathcal{G}}_c^{\sigma} \tag{11}$$

Adversarial Style Perturbation. We get the random initialized global styles $\mathfrak{S}_{init} = \{\boldsymbol{\mu}_{init}, \boldsymbol{\sigma}_{init}\}$ by adding Gaussian noise $\mathcal{N}(0, I)$, where:

$$\boldsymbol{\mu}_{init} = \boldsymbol{\mu}_g + \varepsilon \cdot \mathcal{N}(0, I) \tag{12}$$

$$\boldsymbol{\sigma}_{init} = \boldsymbol{\sigma}_g + \varepsilon \cdot \mathcal{N}(0, I) \tag{13}$$

where ε is set to $\frac{16}{255}$. Then, the ensemble gradients are incorporated into the initialized style to get the adversarial styles $\mathfrak{S}_{adv} = \{\boldsymbol{\mu}_{adv}, \boldsymbol{\sigma}_{adv}\}$, where:

$$\boldsymbol{\mu}_{adv} = \boldsymbol{\mu}_{init} + \kappa_1 \cdot sign(\boldsymbol{\mathcal{G}}_e^{\mu}) \tag{14}$$

$$\boldsymbol{\sigma}_{adv} = \boldsymbol{\sigma}_{init} + \kappa_2 \cdot sign(\boldsymbol{\mathcal{G}}_e^{\sigma}) \tag{15}$$

Notably, κ_1 and κ_2 are chosen randomly from a given set of coefficients, which will not force a consistent change in the degree of the perturbation of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, making the model generate a more diverse range of styles. After obtaining the adversarial styles, style migration is performed with AdaIN method to enhance the generalizability:

$$\boldsymbol{F}_{adv} = \frac{\boldsymbol{F}_g - \boldsymbol{\mu}_g}{\boldsymbol{\sigma}_g} * \boldsymbol{\sigma}_{adv} + \boldsymbol{\mu}_{adv} \tag{16}$$

Then, the adversarial and global feature maps will together be passed to the remainder of the backbone and the FSL classifier to accomplish the N-way K-shot FSL, resulting in two predictions $\boldsymbol{p}_g^{fsl} \in \mathbb{R}^{B\times N_c}$ and $\boldsymbol{p}_{adv}^{fsl} \in \mathbb{R}^{NM\times N}$. Furthermore, we can get $\mathcal{L}_{fsl}$:

$$\mathcal{L}_{fsl} = \mathcal{L}_{CE}(\boldsymbol{p}_g^{fsl}, y_{fsl}) + \mathcal{L}_{CE}(\boldsymbol{p}_{adv}^{fsl}, y_{fsl}) \tag{17}$$

where $y_{fsl} \in \mathbb{R}^{NM}$ is the query samples' logical labels.

DCO. We design a novel objective function named Discrepancy & Consistency Optimization (DCO) to maximize seen-unseen domain visual discrepancy and global-crop consistency for overall features $\mathbb{F}_{all} = \{\boldsymbol{F}_1, \boldsymbol{F}_2, \cdots, \boldsymbol{F}_k, \boldsymbol{F}_g, \boldsymbol{F}_{adv}\}$. For seen-unseen domain discrepancy maximum, we consider the global and crop features to belong to the seen domain and the generated adversarial features to belong to the unseen domain. Therefore, it is possible to make the generated adversarial features located as far away from the source domain as possible. The domain discriminator contains a dropout layer and a fully connected layer. The domain discrepancy loss is:

$$\mathcal{L}_{dom} = \sum\nolimits_{\boldsymbol{F}\in\mathbb{F}_{all}} \mathcal{L}_{CE}(f_{dom}(\boldsymbol{F};\theta_{dom}), d_F) \tag{18}$$

where $d_F \in \{0, 1\}$ is the domain label with 0 (*resp.*, 1) indicating $\boldsymbol{F}$ is from the seen(*resp.*, the unseen) domain. Moreover, we enforce the semantic consistency between the global and crop features as:

$$\mathcal{L}_{con} = \sum_{i=1}^{k}(\lambda\mathcal{L}_{CE}(\boldsymbol{p}_i, \boldsymbol{p}_g) + (1-\lambda)\mathcal{L}_{CE}(\boldsymbol{p}_i^{fsl}, y_{fsl})) \tag{19}$$

where, $\boldsymbol{p}_i^{fsl} = f_{re}(\boldsymbol{F}_i;\theta_{re})$. We use Kullback-Leibler divergence loss $KL(\cdot)$ to maximize global-adversarial consistency as:

$$\mathcal{L}_{adv} = KL(\boldsymbol{p}_{adv}^{fsl}, \boldsymbol{p}_g^{fsl}) \tag{20}$$

Then the final objective loss of **SVasP** is:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{fsl} + \mathcal{L}_{dom} + \mathcal{L}_{con} + \mathcal{L}_{adv} \tag{21}$$

More construction details and the complete adversarial style generation pseudo-code can be found in Appendix A.

Experiments

Datasets

Following the BSCD-FSL benchmark proposed in BSCD-FSL (Guo et al. 2020) and the *mini*-CUB benchmark proposed in FWT (Tseng et al. 2020), we use *mini*ImageNet (Ravi and Larochelle 2017) with 64 classes as the source domain. The target domains include eight datasets: ChestX (Wang et al. 2017), ISIC (Tschandl, Rosendahl, and Kittler 2018), EuroSAT (Helber et al. 2019), CropDisease (Mohanty, Hughes, and Salathé 2016), CUB (Wah et al. 2011), Cars (Krause et al. 2013), Places (Zhou et al. 2017), and Plantae (Van Horn et al. 2018). In our Single Source CD-FSL setting, target domain datasets are not available during meta-training stage.

	Method	FT	ChestX	ISIC	EuroSAT	CropDisease	CUB	Cars	Places	Plantae	Aver.
1-shot	GNN	✗	22.00±0.46	32.02±0.66	63.69±1.03	64.48±1.08	45.69±0.68	31.79±0.51	53.10±0.80	35.60±0.56	43.55
	FWT	✗	22.04±0.44	31.58±0.67	62.36±1.05	66.36±1.04	47.47±0.75	31.61±0.53	55.77±0.79	35.95±0.58	44.14
	ATA	✗	22.10±0.20	33.21±0.40	61.35±0.50	67.47±0.50	45.00±0.50	33.61±0.40	53.57±0.50	34.42±0.40	43.84
	SET-RCL	✗	22.74±0.20	33.33±0.40	65.53±0.60	68.43±0.50	46.98±0.50	32.84±0.40	56.93±0.50	37.43±0.40	45.53
	StyleAdv	✗	22.64±0.35	33.96±0.57	70.94±0.82	74.13±0.78	48.49±0.72	34.64±0.57	58.58±0.83	41.13±0.67	48.06
	SVasP	✗	**23.23±0.35**	**37.63±0.58**	**72.30±0.82**	**75.87±0.73**	**49.49±0.72**	**35.27±0.57**	**59.07±0.81**	**41.22±0.62**	**49.26**
	ATA	✓	22.15±0.20	34.94±0.40	68.62±0.50	75.41±0.50	46.23±0.50	37.15±0.40	54.18±0.50	37.38±0.40	47.01
	StyleAdv	✓	22.64±0.35	35.76±0.52	**72.92±0.75**	**80.69±0.28**	48.49±0.72	35.09±0.55	58.58±0.83	41.13±0.67	49.41
	SVasP	✓	**23.23±0.35**	**37.63±0.63**	72.30±0.83	77.45±0.68	**49.49±0.72**	**38.18±0.61**	**59.07±0.81**	**41.22±0.62**	**49.82**

	Method	FT	ChestX	ISIC	EuroSAT	CropDisease	CUB	Cars	Places	Plantae	Aver.
5-shot	GNN	✗	25.27±0.46	43.94±0.67	83.64±0.77	87.96±0.67	62.25±0.65	44.28±0.63	70.84±0.65	52.53±0.59	58.84
	FWT	✗	25.18±0.45	43.17±0.70	83.01±0.79	87.11±0.67	66.98±0.68	44.90±0.64	73.94±0.67	53.85±0.62	59.77
	ATA	✗	24.32±0.40	44.91±0.40	83.75±0.40	90.59±0.30	66.22±0.50	49.14±0.40	75.48±0.40	52.69±0.40	60.89
	SET-RCL	✗	25.65±0.20	44.93±0.40	83.84±0.40	88.11±0.30	68.05±0.50	47.95±0.40	76.23±0.40	54.70±0.40	61.18
	StyleAdv	✗	26.07±0.37	45.77±0.51	86.58±0.54	93.65±0.39	68.72±0.67	50.13±0.68	77.73±0.62	**61.52±0.68**	63.77
	SVasP	✗	**26.87±0.38**	**51.10±0.58**	**88.72±0.52**	**94.52±0.33**	**68.95±0.66**	**52.13±0.66**	**77.78±0.62**	60.63±0.64	**65.09**
	Fine-tune	✓	25.97±0.41	48.11±0.64	79.08±0.61	89.25±0.51	64.14±0.77	52.08±0.74	70.06±0.74	59.27±0.70	61.00
	BSR	✓	26.84±0.44	54.42±0.66	80.89±0.61	92.17±0.45	69.38±0.76	57.49±0.72	71.09±0.68	61.07±0.76	64.17
	ATA	✓	25.08±0.20	49.79±0.40	89.64±0.30	95.44±0.20	69.83±0.50	54.28±0.50	76.64±0.40	58.08±0.40	64.85
	NSAE	✓	27.10±0.44	54.05±0.63	83.96±0.57	93.14±0.47	68.51±0.76	54.91±0.74	71.02±0.72	59.55±0.74	64.03
	RDC	✓	25.48±0.20	49.06±0.30	84.67±0.30	93.55±0.30	67.77±0.40	53.75±0.50	74.65±0.40	60.63±0.40	63.70
	StyleAdv	✓	26.24±0.35	53.05±0.54	91.64±0.43	96.51±0.28	70.90±0.63	56.44±0.68	**79.35±0.61**	64.10±0.64	67.28
	SVasP	✓	**27.25±0.39**	**55.43±0.59**	**91.77±0.41**	**96.79±0.26**	**72.06±0.65**	**59.99±0.69**	78.91±0.65	**64.21±0.66**	**68.30**

Table 1: Quantitative comparison to state-of-the-arts methods on eight target datasets based on ResNet-10, which is pretrained on *mini*ImageNet. Accuracy of 5-way 1-shot/5-shot tasks with 95 confidence interval are reported. "**FT**" with ✓means finetuning is used, vice versa. "**Aver.**" means "Average Accuracy" of the eight datasets. The optimal results are marked in **bold**.

Implementation Details

Using ResNet-10 (He et al. 2016) as the backbone and GNN as the N-way K-shot classifier, the network is meta-trained for 200 epochs with 120 episodes per epoch. ResNet-10 is pretrained *mini*ImageNet using traditional batch training. The optimizer is Adam with a learning rate of 0.001. Additionally, using ViT-small (Dosovitskiy et al. 2020) as the feature extractor and ProtoNet (Laenen and Bertinetto 2021) as the N-way K-shot classifier, the network is meta-trained for 20 epochs with 2000 episodes per epoch. The optimizer is SGD with a learning rate of 5e-5 and 0.001 for E and f_{re}, respectively. ViT-small is pretrained on ImageNet1K by DINO (Caron et al. 2021). We evaluate the proposed framework during testing by average classification accuracy over 1000 episodes with a 95% confidence interval. Each class contains 5 support samples and 15 query samples. Hyperparameters are set as follows: $\xi = 0.1$, $k = 2$, $\lambda = 0.2$ and choose κ_1, κ_2 from $[0.008, 0.08, 0.8]$. The probability to perform style change is set to 0.2. The details of the finetuning are attached in Appendix A. All the experiments are conducted on a single NVIDIA GeForce RTX 3090.

Experimental Results

Comparison to SOTA methods on ResNet-10. We compare the proposed SVasP with state of the art methods with ResNet-10 as the backbone in Table 1. For a fair comparison, all the competing methods follow the single source training scheme, which is more realistic and difficult. Concretly, nine representative single source CD-FSL methods are introduced including GNN (Garcia and Bruna 2018), FWT (Tseng et al. 2020), ATA (Wang and Deng 2021), SET-RCL (Zhang et al. 2022), StyleAdv (Fu et al. 2023), Fine-tune (Guo et al. 2020), BSR (Liu et al. 2020), NSAE (Liang et al. 2021) and RDC (Li et al. 2022a). As shown, under whether setting, our method outperforms the second-best approach in terms of average accuracy with a clear margin and builds a new state of the art in the majority of domains. More precisely, under 1-shot setting on ResNet-10, SVasP performs better in all domains and surpasses the strongest competitor StyleAdv significantly by $+0.59\%$, $+3.67\%$, $+1.36\%$, $+1.74\%$, $+1.00\%$, $+0.63\%$ on ChestX, ISIC, EuroSAT, CropDisease, CUB, Cars, respectively. Under 5-shot setting on ResNet-10, SVasP performs better in 7 out of 8 domains,

	Method	FT	ChestX	ISIC	EuroSAT	CropDisease	CUB	Cars	Places	Plantae	Aver.
1-shot	StyleAdv	✗	**22.92±0.32**	33.05±0.44	72.15±0.65	**81.22±0.61**	84.01±0.58	40.48±0.57	72.64±0.67	55.52±0.66	57.75
	SVasP	✗	22.68±0.30	**34.49±0.46**	**72.50±0.62**	80.82±0.62	**85.56±0.57**	**40.51±0.59**	**75.93±0.66**	**56.25±0.65**	**58.59**
	PMF	✓	21.73±0.30	30.36±0.36	70.74±0.63	80.79±0.62	78.13±0.66	37.24±0.57	71.11±0.71	53.60±0.66	55.46
	StyleAdv	✓	**22.92±0.32**	33.99±0.46	74.93±0.58	**84.11±0.57**	84.01±0.58	40.48±0.57	72.64±0.67	55.52±0.66	58.57
	SVasP	✓	22.68±0.30	**34.49±0.46**	**75.51±0.57**	83.98±0.55	**85.56±0.57**	**40.51±0.59**	**75.93±0.66**	**56.25±0.65**	**59.36**
	Method	**FT**	**ChestX**	**ISIC**	**EuroSAT**	**CropDisease**	**CUB**	**Cars**	**Places**	**Plantae**	**Aver.**
5-shot	StyleAdv	✗	**26.97±0.33**	47.73±0.44	88.57±0.34	**94.85±0.31**	95.82±0.27	61.73±0.62	88.33±0.40	75.55±0.54	72.44
	SVasP	✗	26.77±0.34	**49.75±0.46**	**88.69±0.35**	93.25±0.36	**95.95±0.23**	**62.60±0.61**	**89.19±0.39**	**76.49±0.50**	**72.84**
	PMF	✓	27.27	50.12	85.98	92.96	-	-	-	-	-
	StyleAdv	✓	**26.97±0.33**	51.23±0.51	90.12±0.33	95.99±0.27	95.82±0.27	66.02±0.64	88.33±0.40	78.01±0.54	74.06
	SVasP	✓	26.77±0.34	**51.62±0.50**	**90.55±0.34**	**96.17±0.30**	**95.95±0.23**	**66.47±0.62**	**89.19±0.39**	**78.67±0.52**	**74.42**

Table 2: Quantitative comparison to state-of-the-arts methods on eight target datasets based on ViT-small, which is pretrained on ImageNet1K by DINO. Accuracy of 5-way 1-shot/5-shot tasks with 95 confidence interval are reported.

Method	SV	$\mathcal{L}_{dom}$	$\mathcal{L}_{con}$	Aver. (%)
Baseline	-	-	-	62.07
Proposed	✓			62.61
	✓	✓		63.69
	✓		✓	64.05
	✓	✓	✓	**65.09**

Table 3: Ablation study of the proposed method with different component combinations. "SV" indicates SV Gradient Ensemble module.

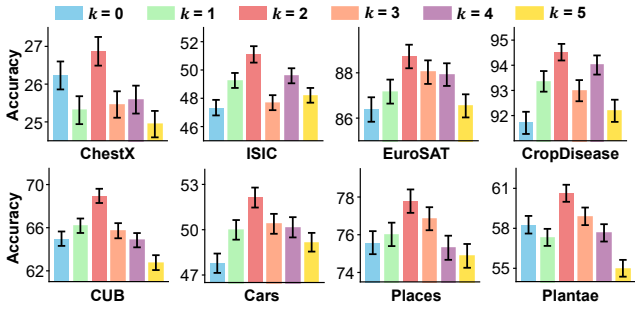

Figure 3: Performances on different numbers of crops k.

and the superiority of SVasP is even larger with higher accuracy by $+0.80\%$, $+5.33\%$, $+2.14\%$, $+0.87\%$, $+2.00\%$ on ChestX, ISIC, EuroSAT, CropDisease, Cars, respectively. Despite being trained on one dataset, SVasP has good generalization ability, thus producing the optimal style-based augmentation policies for the unseen target domains.

Comparison to SOTA methods on ViT-small. To further evaluate the effectiveness of our proposed technique, we apply the proposed SVasP idea to ViT models and compare their performance over other methods on the eight datasets with ViT-small as the backbone and ProtoNet as the classifier. As shown in Table 2, our SVasP is compared with methods like StyleAdv and PMF. SVasP achieves 58.59% and 72.84% top 1 average accuracy on either 5-way 1-shot or 5-way 5shot setting, which outperforms StyleAdv by 0.84%, 0.40%, respectively.

Qualitative Evaluation

We have performed an exhaustive and fair experimental analysis of the proposed method SVasP and the experimental results with ResNet-10 as the backbone and GNN as the classifier under the 5-way 5-shot setting are reported. More experimental results are attached in Appendix B.

Impact of different component in SVasP. To investigate the contribution of different components, we perform an ablation study on SVasP and report the result of average accuracy on eight target domains in Table 3. Specifically, we study the main technical contributions by (a) whether using SV (means the SV Gradient Ensemble module), (b) whether $\mathcal{L}_{dom}$ and (c) whether $\mathcal{L}_{con}$. Among these variants, we can find that the SV Gradient Ensemble module effectively utilizes the source domain style gradients to alleviate the domain shift problem. With well constrained $\mathcal{L}_{dom}$ and $\mathcal{L}_{con}$, SVasP improves the generalization performance and substantially improves the accuracy up to 3.02% on average.

Impact of different crop numbers k. We investigate the optimal solution for the number of crops and find that the model is most robust when the number is set to 2, as illustrated in Figure 3. Because an insufficient number of crops (*e.g.*, 0, 1) fail to represent the style gradients of the source domain and stabilize the global style perturbation. Moreover, excessive crops (*e.g.*, 3, 4, 5) can lead to overfitting of the model and limited by the source domain style.

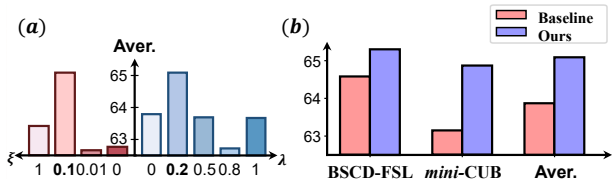

Figure 4: Performances on (a) different ξ, λ and (b) whether use same κ_1, κ_2. The average accuracy (%) is reported.

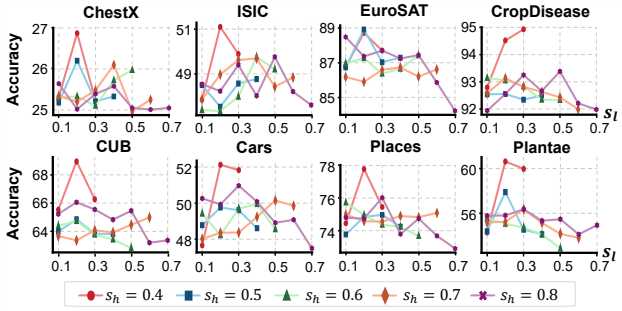

Figure 5: Performances on different scale parameters s.

Impact of different strategies for ξ and λ. The decay factor ξ controls the proportion of the crop style gradients that are incorporated into the global style gradients. In addition, the main component crop consistency loss $\mathcal{L}_{con}$ has a large impact on the performance of the model, which consists of the global-crop prediction consistency loss and the crop FSL classification loss. Performances on different λ and ξ are illustrated in Figure 4 (a). As shown, the accuracy rises as ξ increase from 0 to 0.1, as the proportion increases and provides more source domain gradients. However, the accuracy decreases when the proportion is 1, as too high a proportion of the crop style gradients leads to weak global style gradients. For λ, setting the value of λ to 0.2 can realize an increase in the mean classification accuracy compared to other settings of approximately 1.62%.

Impact of different selection methods for κ_1 and κ_2. Unlike styleadv, which sets κ_1 and κ_2 to the same value, we allow κ_1 and κ_2 to have different values to diversify the style. The experimental results verify the rationality of our setup, as shown in Figure 4 (b).

Impact of different scale parameters s. We evaluate the impact of different scale parameters s, which determines the the area of the crop images. It's important to study optimal values of s because when the area is large, the model overlooks the local regions of inputs. We investigate the performances with $s_l \in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7]$ and $s_h \in [0.4, 0.5, 0.6, 0.7, 0.8]$, with $s_l < s_h$. The optimal result is reached when $s = (0.2, 0.4)$, as shown in Figure 5. We observe that SVasP with smaller area of crop images performs better, which demonstrates the effectiveness of our introduction of localized crop style gradients.

Visualization Results. We visualize the loss landscape following (Li et al. 2018) on the BSCD-FSL bechmark to verify the validity of our proposed important module, as shown in Figure 6. SVasP achieves a stronger flatness which can stand for the better generalization. In additon, in order to provide a more intuitive comparison about the performance of "SVasP"(ours) and "StyleAdv" model, we visualize the class-activation map using the Grad-CAM (Selvaraju et al. 2017) on the BSCD-FSL benchmark, as shown in Figure 7. We can observe that, StyleAdv may pay attention to insignificant things and is disorganized. In contrast, SVasP can focus on more key areas of the target images with the help of the localized crop style gradients. Visualization results on the *mini*-CUB benchmark can be found in Appendix C.

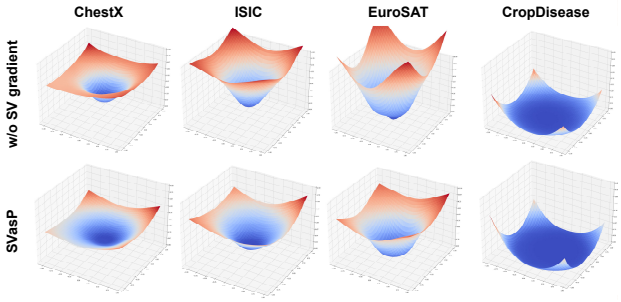

Figure 6: Loss landscape visualization results of the model without SV gradient ensemble module (first row) and our SVasP model (second row) on the BSCD-FSL benchmark.

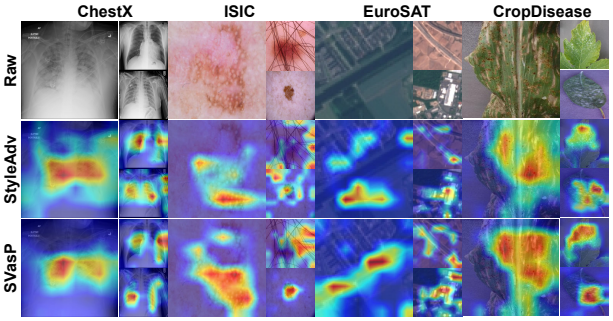

Figure 7: Grad-CAM visualization results of the StyleAdv model and our SVasP model on the BSCD-FSL benchmark. For each target dataset, three examples are demonstrated.

Conclusion

We explore the Single Source Cross-Domain Few-Shot Learning, focusing on the limitations of style-based approaches and addressing the domain shift problem. Our study introduces a novel network to capitalize on the localized crop style gradients, achieving state-of-the-art performance on both ResNet-10 and ViT-small backbone. To enhance the training process, we employ a random cropping strategy and integrate crop style gradients as the style perturbation stabilizers. This approach prevents the model from being confined to the source domain style and local loss minima. Extensive experimental results demonstrate the effectiveness and insights of the proposed method, highlighting its rationality and potential for broader application.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62476056, 62076062, and 62306070) and the Social Development Science and Technology Project of Jiangsu Province (No. BE2022811). Furthermore, the work was also supported by the Big Data Computing Center of Southeast University.

References

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9650–9660.

Chen, C.; Tang, L.; Liu, F.; Zhao, G.; Huang, Y.; and Yu, Y. 2022. Mix and reason: Reasoning over semantic topology with data mixing for domain generalization. In *Proceedings of Advances in Neural Information Processing Systems*, 33302–33315.

Chen, W.-Y.; Liu, Y.-C.; Kira, Z.; Wang, Y.-C. F.; and Huang, J.-B. 2018a. A Closer Look at Few-shot Classification. In *Proceedings of International Conference on Learning Representations*.

Chen, Z.; Badrinarayanan, V.; Lee, C.-Y.; and Rabinovich, A. 2018b. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of International Conference on Machine Learning*, 794–803.

Das, D.; Yun, S.; and Porikli, F. 2021. ConfeSS: A framework for single source cross-domain few-shot learning. In *Proceedings of the International Conference on Learning Representations*, 1–12.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations*, 1–12.

Du, Z.; Li, J.; Su, H.; Zhu, L.; and Lu, K. 2021. Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3937–3946.

Feng, F.; Wang, J.; and Geng, X. 2024. Transferring Core Knowledge via Learngenes. *arXiv preprint arXiv:2401.08139*.

Feng, F.; Wang, J.; Zhang, C.; Li, W.; Yang, X.; and Geng, X. 2023. Genes in Intelligent Agents. *arXiv preprint arXiv:2306.10225*.

Feng, F.; Xie, Y.; Wang, J.; and Geng, X. 2024a. Redefining ¡Creative¿ in Dictionary: Towards a Enhanced Semantic Understanding of Creative Generation. *arXiv preprint arXiv:2410.24160*.

Feng, F.; Xie, Y.; Wang, J.; and Geng, X. 2024b. Wave: Weight template for adaptive initialization of variable-sized models. *arXiv preprint arXiv:2406.17503*.

Fu, Y.; Fu, Y.; and Jiang, Y.-G. 2021. Meta-fdmixup: Cross-domain few-shot learning guided by labeled target data. In *Proceedings of the ACM International Conference on Multimedia*, 5326–5334.

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*, 24575–24584.

Garcia, V.; and Bruna, J. 2018. Few-shot learning with graph neural networks. In *Proceedings of the International Conference on Learning Representations*, 1–12.

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 European Conference on Computer Vision*, 124–141.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

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, Y.; and Ma, A. J. 2022. Adversarial feature augmentation for cross-domain few-shot classification. In *Proceedings of the European Conference on Computer Vision*, 20–37.

Islam, A.; Chen, C.-F. R.; Panda, R.; Karlinsky, L.; Feris, R.; and Radke, R. J. 2021. Dynamic distillation network for cross-domain few-shot recognition with unlabeled data. In *Proceedings of Advances in Neural Information Processing Systems*, 3584–3595.

Kim, T.; and Han, B. 2024. Randomized adversarial style perturbations for domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2317–2325.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 554–561.

Laenen, S.; and Bertinetto, L. 2021. On episodes, prototypical networks, and few-shot learning. In *Proceedings of Advances in Neural Information Processing Systems*, 24581–24592.

Li, H.; Xu, Z.; Taylor, G.; Studer, C.; and Goldstein, T. 2018. Visualizing the loss landscape of neural nets. In *Proceedings of Advances in Neural Information Processing Systems*, 1–11.

Li, L.; Xiao, J.; Chen, G.; Shao, J.; Zhuang, Y.; and Chen, L. 2024. Zero-shot visual relation detection via composite visual cues from large language models. In *Proceedings of Advances in Neural Information Processing Systems*.

Li, P.; Gong, S.; Wang, C.; and Fu, Y. 2022a. Ranking distance calibration for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9099–9108.

Li, X.; Dai, Y.; Ge, Y.; Liu, J.; Shan, Y.; and DUAN, L. 2022b. Uncertainty Modeling for Out-of-Distribution Generalization. In *International Conference on Learning Representations*, 1–13.

Liang, H.; Zhang, Q.; Dai, P.; and Lu, J. 2021. Boosting the generalization capability in cross-domain few-shot learning via noise-enhanced supervised autoencoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9424–9434.

Liu, B.; Zhao, Z.; Li, Z.; Jiang, J.; Guo, Y.; and Ye, J. 2020. Feature transformation ensemble model with batch spectral regularization for cross-domain few-shot classification. *arXiv preprint arXiv:2005.08463*.

Mohanty, S. P.; Hughes, D. P.; and Salathé, M. 2016. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7: 215232.

Qi, L.; Dong, P.; Xiong, T.; Xue, H.; and Geng, X. 2024. DoubleAUG: Single-domain Generalized Object Detector in Urban via Color Perturbation and Dual-style Memory. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(5): 1–20.

Rame, A.; Dancette, C.; and Cord, M. 2022. Fishr: Invariant gradient variances for out-of-distribution generalization. In *Proceedings of the International Conference on Machine Learning*, 18347–18377.

Ravi, S.; and Larochelle, H. 2017. Optimization as a model for few-shot learning. In *Proceedings of the International Conference on Learning Representations*, 1–11.

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/CVF International Conference on Computer Vision*, 618–626.

Triantafillou, E.; Zhu, T.; Dumoulin, V.; Lamblin, P.; Evci, U.; Xu, K.; Goroshin, R.; Gelada, C.; Swersky, K.; Manzagol, P.-A.; et al. 2020. Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples. In *International Conference on Learning Representations*, 1–13.

Tschandl, P.; Rosendahl, C.; and Kittler, H. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1): 1–9.

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*, 1–14.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8769–8778.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, H.; and Deng, Z.-H. 2021. Cross-domain few-shot classification via adversarial task augmentation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1075–1081.

Wang, X.; and He, K. 2021. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1924–1933.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2097–2106.

Xie, Y.; Feng, F.; Wang, J.; Geng, X.; and Rui, Y. 2024. Kind: Knowledge integration and diversion in diffusion models. *arXiv preprint arXiv:2408.07337*.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient surgery for multi-task learning. In *Proceedings of Advances in Neural Information Processing Systems*, 5824–5836.

Zhang, J.; Song, J.; Gao, L.; and Shen, H. 2022. Free-lunch for cross-domain few-shot learning: Style-aware episodic training with robust contrastive learning. In *Proceedings of the ACM International Conference on Multimedia*, 2586–2594.

Zhang, X.; Xu, R.; Yu, H.; Zou, H.; and Cui, P. 2023. Gradient norm aware minimization seeks first-order flatness and improves generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20247–20257.

ZHENG, H.; Wang, R.; Liu, J.; and Kanezaki, A. 2023. Cross-Level Distillation and Feature Denoising for Cross-Domain Few-Shot Classification. In *Proceedings of the International Conference on Learning Representations*, 1–12.

Zhong, Z.; Zhao, Y.; Lee, G. H.; and Sebe, N. 2022. Adversarial style augmentation for domain generalized urban-scene segmentation. *Advances in Neural Information Processing Systems*, 35: 338–350.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6): 1452–1464.

Zhou, F.; Wang, P.; Zhang, L.; Wei, W.; and Zhang, Y. 2023. Revisiting prototypical network for cross domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20061–20070.

Zhou, K.; Yang, Y.; Qiao, Y.; and Xiang, T. 2020. Domain Generalization with MixStyle. In *Proceedings of the International Conference on Learning Representations*, 1–12.

Zhuo, L.; Fu, Y.; Chen, J.; Cao, Y.; and Jiang, Y.-G. 2022. Tgdm: Target guided dynamic mixup for cross-domain few-shot learning. In *Proceedings of the ACM International Conference on Multimedia*, 6368–6376.

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*, 23575–23584.