D-GAP: Improving Out-of-Domain Robustness via Dataset-Agnostic and Gradient-Guided Augmentation in Amplitude and Pixel Spaces

Ruoqi Wang
Institute of AI for Industries, Chinese Academy of Sciences
The Hong Kong University of Science and Technology (Guangzhou)
rqwang@iaii.ac.cn

Haitao Wang
Sun Yat-sen University
wanght76@mail.sysu.edu.cn

Shaojie Guo
East China Normal University
52275904013@stu.ecnu.edu.cn

Qiong Luo*
The Hong Kong University of Science and Technology
The Hong Kong University of Science and Technology (Guangzhou)
luo@cse.ust.hk

Abstract

Out-of-domain (OOD) robustness is challenging to achieve in real-world computer vision, especially in unsupervised domain adaptation scenarios, where shifts in image background, style, and acquisition instruments often degrade model performance. Generic augmentations show inconsistent gains under such shifts, whereas dataset-specific augmentations require expert knowledge and prior analysis. Moreover, prior studies show that neural networks adapt poorly to domain shifts because they exhibit a learning bias to domain-specific frequency components. Perturbing frequency values can mitigate such bias but overlooks pixel-level details, leading to suboptimal performance. To address these limitations, we propose **D-GAP**, a **D**ataset-agnostic and **G**radient-guided augmentation method for the **A**mplitude spectrum (in frequency space) and the **P**ixel values. Unlike conventional handcrafted augmentations, D-GAP computes sensitivity maps in the frequency space from task gradients, which reflect how strongly the deep models respond to different frequency components, and uses the maps to adaptively interpolate amplitudes between source and target samples. We further propose a dual-space augmentation that jointly controls spectral bias and spatial fidelity by introducing a complementary pixel-space blending branch. This way, D-GAP turns augmentation from fixed, random, or manually designed perturbation into a model-response-adaptive intervention. Extensive experimental results show that the proposed method consistently outperforms both generic and dataset-specific domain adaptation methods, improving average OOD performance by +5.3% on four real-world datasets and +1.9% on three benchmark datasets. Code is available at https://github.com/RapidsAtHKUST/D-GAP.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

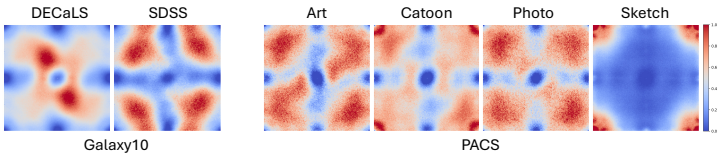


Figure 1: Average amplitude sensitivity maps of ResNet50 trained across domains.

1 Introduction

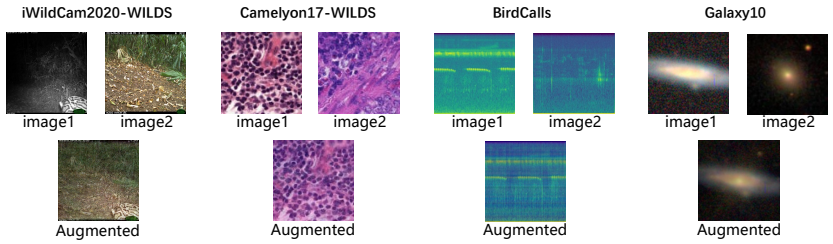
In real-world applications, the distributions in the training datasets often differ from those in other deployment environments because of the shifts in background and acquisition instruments [26, 42]. In such out-of-distribution (OOD) settings, models trained on source domains always degrade when applied to unlabeled target domains. For example, in wildlife monitoring, ecologists use models to classify species of camera trap images, but these models often show significant accuracy drops when applied to new locations [3]. Annotated data are usually available only for a limited subset of cameras, which may not contain the diversity of environments in the unlabeled target domains. In this paper, we focus on improving OOD robustness in domain adaptation settings, where we have labeled data from source domains and unlabeled data from target domains.

Data augmentation is a common strategy for improving OOD robustness of deep-learning models, but designing effective augmentations remains challenging. Generic augmentations of images (e.g., RandAugment [6], CutMix [74], FACT [66], SAM [65]) improves the performance across datasets to various degrees [12, 61, 16], whereas domain-invariance methods [68, 76, 23, 12, 72, 62, 52] yield limited improvements, especially on real-world domain shifts [11]. Although some dataset-specific augmentation methods [11, 42] can improve OOD performance, they often need expert knowledge and prior analysis of the dataset, which makes these methods hard to scale or apply to new datasets. These limitations highlight the need for more general, dataset-agnostic methods to improve OOD robustness across diverse real-world settings.

Prior studies show that neural networks tend to learn biasedly from different frequency components depending on the dataset's frequency characteristics [40, 15, 60]. Such spectral bias causes the model to overfit domain-specific frequency components, leading to poor generalization when the target domain has different frequency distributions. As such, we consider that this bias can be reduced by perturbing the the frequency distributions. However, existing frequency-based augmentation methods [65, 66, 71, 70, 73, 34, 30, 47, 22] typically perturb frequency components using predefined, manually constrained rules, or cannot integrate information from unlabeled target domain. This lack of model-response awareness limits their performance in domain adaptation tasks.

To address these problems, we propose **D-GAP**, a domain adaptation framework that improves OOD robustness by introducing targeted augmentations in both frequency and pixel spaces. Unlike previous targeted augmentations with manually designed, dataset-specific rules, D-GAP automatically adapts to domain shifts from data and models, without the need for expert knowledge or prior analysis. Specifically, given two images from different domains, in the frequency space, we introduce a gradient-adaptive mechanism that computes amplitude sensitivity maps from the task gradients with respect to the amplitude spectrum (as shown in Fig. 1). These sensitivity maps represent how strongly the models respond to different amplitude components. Instead of interpolating two spectra with a predefined or random ratio, D-GAP adjusts the interpolation strength at each frequency according to its sensitivity value: more sensitive amplitude components are more strongly mixed with target-domain amplitude spectrum, while less sensitive ones are more preserved. This mechanism enables the models to suppress domain-specific spectral learning bias. This gradient-guided interpolation creates frequency-space augmentations that adaptively keep the main content of the source image while adding features from target domains. However, reconstructed results of amplitude blending sometimes introduce artifacts and blurring. As such, we apply a pixel value blending to add complementary spatial pixel-level information. The final augmented image is then obtained by blending these two results from both the frequency space and the pixel space.

We firstly evaluate our method on four real-world datasets: Compared to other generic methods, D-GAP achieves improvements in OOD performance, including +2.1% gains on iWildCam [2, 45] (wildlife recognition), +4.2% on Camelyon17 [1, 45] (tumor detection), +5.6% on BirdCalls [24, 11]



	iWildCam2020-WILDS	Camelyon17-WILDS	BirdCalls	Galaxy10
x_{obj}	animal foreground	cell morphology	bird calls	galaxy structure
$x_{d:robust}$	habitat features in background	cancer stage, tumor size and density	habitat noise	viewing angle
$x_{d:spu}$	low-level background features	stain color	microphone gain settings low-level noise	low-level noise background features
x_{noise}	xy-position	patch orientation, cell xy-positions	x-position	xy-position

Figure 2: **Feature decomposition and augmentation examples.** *Top:* For each dataset (iWildCam, Camelyon17, BirdCalls, Galaxy10), we show a source image, its corresponding augmented image generated by our method. *Bottom:* Based on the decomposition framework [49], we annotate representative features in the datasets across x_{obj}, $x_{d:\text{robust}}$, $x_{d:\text{spu}}$, x_{noise}. Our method effectively randomizes $x_{d:\text{spu}}$, varies $x_{d:\text{robust}}$ while preserving or x_{obj}.

(birdcall recognition), and +9.3% on Galaxy10 [18] (galaxy morphology classification). Also, our method surpasses dataset-specific augmentation strategies tailored for different datasets. Additionally, D-GAP also makes average gains of +1.9% accuracy on standard benchmark datasets (PACS [31], Office-Home [57], and Digits-DG [76]), showing its improvement across diverse distribution shifts. We further analyze the generalizability of D-GAP on various backbone networks and its effectiveness of connecting same-class cross-domain samples while separating different-class examples.

Our work makes the following contributions: (1) We propose a model-response-adaptive augmentation for OOD robustness in domain adaptation, where the perturbation intensity is determined by the amplitude sensitivity of models rather than by manually designed rules. (2) We propose Gradient-guided Amplitude Mix, which computes gradient-guided amplitude sensitivity maps and performs cross-domain interpolation. (3) We introduce a dual-space augmentation design that couples gradient-guided spectral intervention with pixel-space detail compensation. (4) We achieve SOTA results across multiple backbones on both real-world and general benchmark datasets.

2 Background and Formulation

2.1 Feature Decomposition

To better understand how models perform under domain shifts, we use the feature decomposition framework [11], which categorizes input features based on their dependence on the label and domain. Specifically, input features can be decomposed into four types based on their dependence on the label and the domain: (1) features that are *label-dependent* and *domain-independent* (x_{obj}); (2) *label-dependent* and *domain-dependent* features ($x_{\text{d:robust}}$); (3) *label-independent* and *domain-dependent* features ($x_{\text{d:spu}}$); and (4) *label-independent* and *domain-independent* features (x_{noise}). We formalize these relationships via their dependence with the label y and domain d:

$$x_{\text{obj}},\ x_{\text{d:robust}} \not\perp\!\!\!\perp\ y, \quad x_{\text{noise}},\ x_{\text{d:spu}} \perp\!\!\!\perp\ y, \quad x_{\text{d:robust}},\ x_{\text{d:spu}} \not\perp\!\!\!\perp\ d, \quad x_{\text{obj}},\ x_{\text{noise}} \perp\!\!\!\perp\ d.$$

Models should focus on both x_{obj} and $x_{\text{d:robust}}$ while avoiding reliance on spurious or noisy features [11]. We later analyse these relationships using a connectivity-based framework (Section 2.4 and 4.6) to evaluate how augmentations influence these label/domain dependent features.

2.2 Datasets

Real-World Datasets. As shown in Fig. 2, we test our method on four real-world datasets in this paper: **iWildCam** [2, 26] – wildlife species classification with location-related background shifts. **Camelyon17** [1, 26] – tumor identification across hospitals with stain and density variations.

BirdCalls [24, 11] – bird species recognition under habitat and microphone domain noise. **Galaxy10** [18] – galaxy morphology classification across telescopes differing in resolution and color profile. Each dataset presents distinct domain shifts such as environmental background, staining style, recording device, or instrument differences. More dataset details are provided in Appendix Sec D.

Additional Benchmark Datasets. In addition to the four real-world datasets used throughout this paper, we further evaluate our method on three standard benchmarks: **PACS** [31], **Office-Home** [57], and **Digits-DG** [76]. Compared to our four real-world datasets that cover diverse real-world scenario and deployment-related shifts, these benchmark datasets focus on visual recognition with well-studied style variations. More details are provided in Appendix Sec. D.

2.3 Targeted Augmentations

Some dataset-specific augmentation methods [11, 42] can improve OOD performance on real-world applications. However, they often need expert knowledge and prior analysis of the dataset, which makes these methods hard to scale or apply to new datasets. For instance, in wildlife recognition, some methods randomize the background of camera trap images to mitigate spurious correlations between results and location-specific terrains or lighting conditions. Such a target augmentation helps the model to focus on the animal foreground. However, it depends on the segmentation of foreground and background, restricting its application to datasets without foreground annotations. Similarly, in tumor identification from histopathology slides, Stain Color Jitter can deal with stain color variations across hospitals [56]. It perturbs images within the hematoxylin and eosin staining features to overcome color-related domain biases, while preserving morphological features [11]. However, it requires expert knowledge of staining chemistry and color editing. These examples show that while dataset-specific augmentations can be effective, they are difficult to generalize.

2.4 Connectivity

To better understand how data augmentations impact cross-domain generalization, we adopt the concept of *connectivity* [49]. We define connectivity between a class-domain pair $((y_1, d_1), (y_2, d_2))$ under four conditions: (1) $\rho : y_1 = y_2,\ d_1 = d_2$. (2) $\alpha : y_1 = y_2,\ d_1 \neq d_2$. (3) $\beta : y_1 \neq y_2,\ d_1 = d_2$. (4) $\gamma : y_1 \neq y_2,\ d_1 \neq d_2$. Here, each of the $\rho, \alpha, \beta, \gamma$ values indicates how "connected" a pair is. Shen et al. [49] show that the ratios $\frac{\alpha}{\gamma}$ and $\frac{\beta}{\gamma}$ exhibit strong empirical correlation with OOD accuracy. In this work, we follow this approach to show how our augmentation changes the connectivity of the data. More details on the computation of connectivity are presented in Section 4.6 and the Appendix Sec. F. To improve OOD robustness, the model must randomize features that depend on specific domains while preserving task-relevant information. Specifically, in feature decomposition [49], we randomize domain-dependent features ($x_{d:\text{spu}}$) and keep label-relevant ones (x_{obj}, $x_{d:\text{robust}}$). As shown in Fig. 3, we observe that amplitude mixing in frequency domain (Details of the mixing strategy are provided in Section 3.) preserves task-relevant semantic structure [66] (x_{obj} and $x_{d:\text{robust}}$), while effectively randomizing $x_{d:\text{spu}}$. However, solely frequency mixing may introduce artifacts and blurring (as examples in Fig. 3 and 4), influencing the extraction of detailed features.

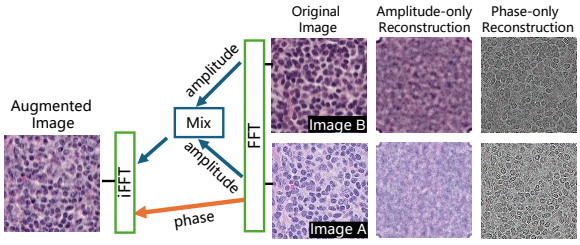


Figure 3: Amplitude-only and phase-only reconstructions of two images. The amplitudes of Image A and B are first mixed, and the mixed amplitude is combined with the phase of Image A to generate an augmentation.

To address the problem, we add pixel-space mixing as a complement, which brings back more pixel-level details and perturbs $x_{d:\text{spu}}$. However, it may influence x_{obj} and $x_{d:\text{robust}}$ as well. Therefore, we then define a tunable ratio (details are in Section 3). By combining the augmentation in both frequency and pixel spaces, our goal is to balance the label-relevant and domain-relevant shifts, achieving domain diversity

Table 1: Connectivity values on iWildCam.

Aug	α	β	γ	α/γ	β/γ
w/o	0.007	0.008	0.021	0.33	0.38
Ours	0.212	0.030	0.051	4.16	0.59

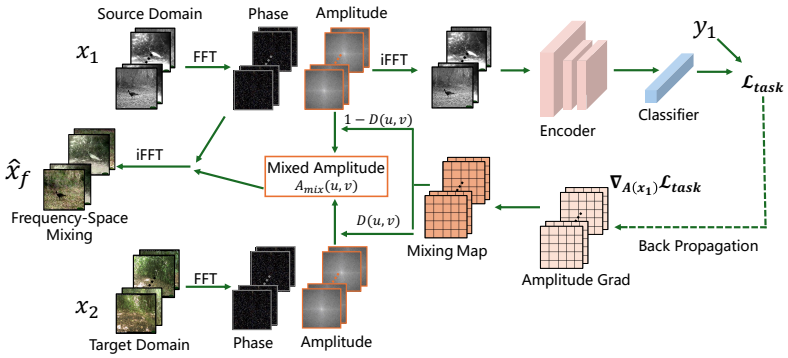


Figure 4: Overview of the Gradient-guided Amplitude Mix procedure. Given a source image x_1 and a target-domain image x_2, we compute the Sensitivity Map $G(u, v)$ and the Mixing Map $D(u, v)$. The mixed amplitude is then combined with the phase of x_1.

while maintaining semantic identity. This dual-space augmentation enables fine-grained control over domain perturbation intensity and promotes robust feature learning across source and unseen domains. We then empirically evaluate our method by comparing the connectivity ratios $\frac{\alpha}{\gamma}$ and $\frac{\beta}{\gamma}$ of the augmented and unaugmented datasets. Results in Table 1 show that augmentation in frequency-pixel spaces can enhance the cross-domain connectivity, achieving domain diversity while maintaining label-relevant identity.

3 Method

In this section, we introduce the technical details of D-GAP, including Fourier transform preliminaries and details of each module in our method, including the Gradient-guided Amplitude Mix, Pixel-Space Mixing, and the training framework.

3.1 D-GAP

Fourier Transform of Images. For an image x, its Fourier transformation $\mathcal{F}(x)$ is formulated as:

$$\mathcal{F}(x)(u,v) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} x(h,w)e^{-j2\pi\left(\frac{hu}{H}+\frac{wv}{W}\right)}, \tag{1}$$

where (u, v) are coordinates in the frequency space. $\mathcal{F}^{-1}(x)$ denotes the corresponding inverse Fourier transformation. Both forward and inverse Fourier transforms can be efficiently computed via the Fast Fourier Transform (FFT) algorithm.

The amplitude and phase components of $\mathcal{F}(x)$ are computed respectively as:

$$\mathcal{A}(x)(u,v) = \left[R^2(x)(u,v) + I^2(x)(u,v)\right]^{1/2}, \mathcal{P}(x)(u,v) = \arctan\left(\frac{I(x)(u,v)}{R(x)(u,v)}\right). \tag{2}$$

where $R(x)$ and $I(x)$ denote the real and imaginary parts of $\mathcal{F}(x)$, respectively. For RGB images, the Fourier transform is applied to each channel separately.

Gradient-guided Amplitude Mix. Conventional frequency-based augmentations [66, 65] typically mix amplitudes of two images with a random intensity factor. While such random interpolation introduces diversity, it fails to distinguish how sensitive that the model is to each frequency component. Moreover, the interpolation ratio needs to be manually constrained within a predefined range to avoid excessive distortion. To overcome these limitations, D-GAP performs a gradient-guided amplitude mix, adaptively adjusting the interpolation strength based on the model's sensitivity in the frequency space. The process of Gradient-guided Amplitude Mix is shown in Fig. 4.

Given two images x_1 and x_2, we aim to augment x_1 by injecting the amplitude information of x_2 into a square region Ω_r in the Fourier space. Let f denote the classifier and $\mathcal{L}_{\text{task}}$ the supervised loss.

We compute the gradient of the loss with respect to the source amplitude:

$$G(u,v) = \Big| \frac{\partial\, \mathcal{L}_{\text{task}}(f(x_1), y)}{\partial\, A(x_1)(u,v)} \Big|, \quad (u,v) \in \Omega_r. \tag{3}$$

Here, $G(u,v)$ measures the model's sensitivity to each frequency, which means how much the task loss would change if the amplitude at a specific frequency (u,v) were perturbed. In this paper, we call $G(u,v)$ **sensitivity map**. Large gradient values indicate that the model's prediction heavily depends on that frequency component, implying stronger spectral bias. This allows D-GAP to continuously regulate how much information to mix at each frequency.

Then after normalizing G and applying a sigmoid function,

$$\tilde{G} = \frac{G - \mu(G)}{\sigma(G) + \varepsilon}, \quad D = \text{clip}\big(\text{Sigmoid}(\tilde{G}), d_{\min}, d_{\max}\big), \tag{4}$$

we obtain a gradient-based **mixing map** $D(u,v)$ from the sensitivity map $G(u,v)$ that adaptively controls the amplitude interpolation. In our design, higher sensitivity (larger $G(u,v)$) leads to stronger interpolation from the target domain, encouraging the model to replace its frequency-biased components with cross-domain information. Conversely, D-GAP uses weaker interpolation in lower sensitivity regions to preserve non-biased features. The final adaptive interpolation is computed as:

$$A_{\text{mix}}(u,v) = \big(1 - D(u,v)\big)\, A(x_1)(u,v) + D(u,v)\, A(x_2)(u,v), \qquad (u,v) \in \Omega_r. \tag{5}$$

Finally, the mixed image is reconstructed as

$$\mathcal{F}_{\text{mix}}(u,v) = A_{\text{mix}}(u,v)\, e^{-jP(x_1)(u,v)}, \hat{x}_f = \mathcal{F}^{-1}\big(\mathcal{F}_{\text{mix}}\big). \tag{6}$$

Through this gradient-guided strategy, D-GAP adaptively perturbs spectral values with target domain more strongly where the model exhibits higher frequency bias. This data-driven, continuous modulation mitigates models' reliance on domain-specific spectral shortcuts, encouraging the model to learn more balanced and generalizable representations across domains.

Pixel-Space Mixing: Frequency blending sometimes introduces artifacts and blurring in the resulting augmentations. As a result, in addition to frequency-domain augmentation, we introduce a pixel-space mixing strategy to add more pixel-level details in the augmentation. Specifically, given a source image x and a target image x', we apply pixel-wise blending with ratio λ_1:

$$\hat{x}_p = (1 - \lambda_1)x_1 + \lambda_1 x_2, \tag{7}$$

Then we combine pixel-space and frequency-space augmented images via a second-stage blending:

$$\hat{x} = (1 - \lambda_2)\hat{x}_f + \lambda_2 \hat{x}_p, \tag{8}$$

This two-stage fusion allows us to perturb frequency bias while maintaining the semantic content of the original sample and the pixel details.

3.2 Training Framework

Our training strategy follows established practices in prior works [42, 28, 66, 65]: For **real-world datasets** (e.g., *iWildCam*, *Camelyon17*, *BirdCalls*, and *Galaxy10*), we use the **Linear Probing then Fine-Tuning (LP-FT)** strategy [42, 28]. We first train a linear classifier on frozen pretrained features to stabilize early optimization, then fine-tune both the encoder and classifier using D-GAP augmentations. This procedure improves OOD robustness by gradually adapting high-level representations to domain-diverse augmentations while preventing early-stage overfitting. For other **common DG benchmark datasets** (i.e., *PACS*, *Office-Home*, *Digits-DG*), we follow previous works [66, 65] to train directly on the pretrained encoder without the LP-FT stage for fair comparisons. On these datasets, the pretrained model already provides a stable initialization, and applying D-GAP augmentations directly during training yields efficient adaptation and robust convergence.

4 Experiments

In this section, we first conduct comparisons between D-GAP and other recent methods on various datasets. We also conduct ablation studies to evaluate the contribution of each module in D-GAP. We then evaluate how our method generalizes beyond backbones. Finally, we conduct a connectivity analysis on empirical evidence of how our method enhances cross-domain feature alignment. Additionally, we present the analysis over hyperparameters including Ω_r and the blending weights λ_1, λ_2 in Appendix Sec. B.

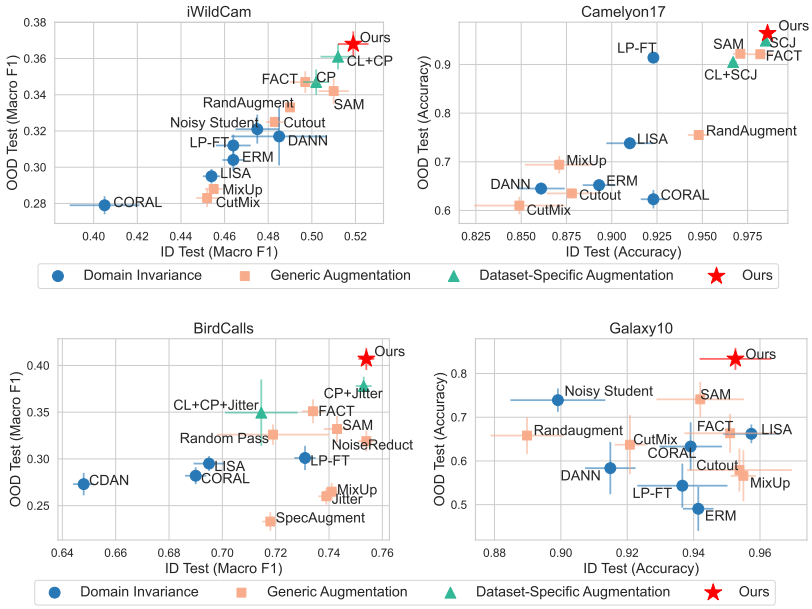


Figure 5: We plot the in-domain (ID) versus OOD performance across datasets (mean and std). 'CL' refers to 'Connect Later'. 'CP' refers to 'Copy Paste'. 'SCJ' refers to 'Stain Color Jitter'.

4.1 Experimental Settings

Platform: We conduct the experiments on a server with 128 AMD EPYC 7543 32-Core Processors, 512GB main memory, and eight NVIDIA RTX A6000 GPUs, each with 48GB device memory. The operating system is Ubuntu 22.04. Our model is implemented in PyTorch 2.1.0.

Evaluation Metrics: Following previous work [11, 42, 48, 58, 66, 65], we select metrics that best suit the characteristics of each dataset to ensure a comprehensive evaluation of model performance. We use F1 score and accuracy as evaluation metrics, depending on the nature of the task and dataset. The F1 score is used for iWildCam, BirdCalls datasets to account for class imbalance, and we use accuracy for Camelyon17, Galaxy10, PACS, Digits-DG and OfficeHome.

Method under Comparison: We compare our method against various baseline approaches for domain generalization and out-of-domain (OOD) performance enhancement. These baselines include:

(1) Generic Augmentations and Domain Invariance Baselines. For **real-world datasets** iWildCam, Camelyon17 and Galaxy10, we compare our approach with several commonly used augmentation techniques: RandAugment [6], CutMix [74], MixUp [75], Cutout [7], FACT [66], and SAM [65]. In the case of the BirdCalls dataset, we compare with MixUp, SpecAugment [39], random low/high pass filters, noise reduction [46], Jitter [56, 11], FACT [66], and SAM [65]. We also compare to LISA [72], a method designed to encourage domain invariance by applying mixing augmentations to inputs of the same class across different domains. Additionally, we evaluate other domain invariance algorithms that do not rely on augmentation, including DANN [33, 10], DeepCORAL [54, 53], and ERM [42], LP-FT [28]. For the **benchmark datasets**, we compared with four SOTA baselines FACT [66], SAM [65], Su et al, [52] (following the setting of DA + DeepCoral), and AGST [62].

(2) Dataset-Specific Targeted Augmentations. We also compare our method with dataset-specific targeted augmentations [11] for real-world datasets: Copy-Paste for iWildCam, Stain Color Jitter for Camelyon17, and Copy-Paste + Jitter for BirdCalls. We also compare to Connect Later, which includes the dataset-specific targeted augmentations in finetuning after pretraining [42].

Backbone Networks: In section 4.2, 4.3, we use the same pretrained backbones as [42, 11] for iWildCam (ResNet50 [14]), Camelyon17 (DenseNet-121 [21]) and BirdCalls (EfficientNet-B0 [55]). For Galaxy10, we follow [59] to use ResNet18 [14] for better performance. And we use the same ConvNet backbone as [76, 65, 77, 66] for Digits-DG, and ResNet18 [14] for OfficeHome and PACS.

We additionally evaluate our method in 4.4 with ViT [8] and ConvNeXt [32] backbone networks to verify that D-GAP generalizes beyond backbone choices.

4.2 Comparison Results

Results on Real-World Datasets. Fig. 5 summarizes the in-domain (ID) and out-of-domain (OOD) performance for all methods across four real-world datasets: iWildCam, Camelyon17, BirdCalls, and Galaxy10. Each point represents one method, colors and shapes represent method types, and our method is marked as a red star. Our method consistently achieves the best OOD performance across all four real-world datasets, while maintaining a strong ID performance. Notably, our approach significantly outperforms all generic methods, including **generic augmentation** and **domain invariance** baselines that do not require dataset-specific design or prior knowledge. Compared to the best generic method (FACT) on iWildCam, our method improves OOD Macro F1 from 34.7% to 36.8%. On Camelyon17, our approach achieves 96.4% OOD accuracy, outperforming the best generic augmentation (SAM at 92.2%) and the best domain invariance method (LP-FT at 91.4%). On BirdCall, OOD Macro F1 increases from 35.1% (FACT) to 40.7%, and on Galaxy10, our method improves OOD accuracy from 74.1% (SAM) to 83.4%.

Moreover, our method also outperforms **dataset-specific augmentation** methods Connect Later + Copy-Paste on iWildCam, Connect Later + Stain Color Jitter on Camelyon17, and Connect Later + Copy-Paste + Jitter on BirdCalls, all of which require expert knowledge and manual design tailored for the dataset based on its characteristics. In contrast, our method is in a fully **dataset-agnostic** way, but still achieves superior or comparable results. These results show that our method is not only effective but also broadly applicable, without relying on dataset-specific analysis or prior knowledge.

Table 2: Model accuracy of leave-one-domain-out evaluation on PACS and OfficeHome.

Methods	PACS					OfficeHome				
	Art	Cartoon	Photo	Sketch	Avg.	Art	Clipart	Product	Real	Avg.
DeepAll	84.94	76.98	97.64	76.75	84.08	57.88	52.72	73.50	74.80	64.72
FACT (2023) [66]	**89.68**	81.06	97.02	83.76	87.88	60.34	54.85	74.48	76.55	66.56
SAM (2023) [65]	85.25	82.02	96.75	85.04	87.27	60.79	**55.47**	74.37	76.37	66.75
Su et al. (2024) [52]	79.88	78.03	95.93	73.15	81.75	64.13	54.83	75.60	75.36	67.71
AGST (2025) [62]	83.50	80.83	96.49	79.57	85.35	62.50	53.87	74.57	75.70	66.66
Ours	89.34	**82.81**	**97.99**	**85.98**	**89.03**	**64.82**	55.21	**77.76**	**83.08**	**70.22**

Table 3: Model accuracy of leave-one-domain-out evaluation on Digits-DG.

Methods	MNIST	MNIST-M	SVHN	SYN	Avg.
DeepAll	95.8	58.8	61.7	78.6	73.7
FACT (2023) [66]	97.9	65.6	72.4	90.3	81.5
SAM (2023) [65]	97.8	66.8	73.2	90.7	82.1
Su et al. (2024) [52]	97.3	65.9	74.5	92.5	82.6
AGST (2025) [62]	97.3	63.1	70.2	93.2	81.0
Ours	**98.3**	**68.5**	**77.2**	**93.8**	**84.5**

Results on Common Benchmarks. Tables 2 and 3 summarize the leave-one-domain-out results on PACS, Office-Home and Digits-DG (Mean). The results of FACT and SAM in these tables are from the source paper [65] for fair comparison, as our re-implementations under identical settings produced slightly lower numbers. DeepAll refers to vanilla backbones trained from source data [67]. In Tables 2 and 3, our method achieves the highest average accuracy on all three benchmarks. These

Table 4: Ablation results across four real-world datasets.

	iWildCam (F1)		Camelyon17 (Acc)		BirdCall (F1)		Galaxy10 (Acc)	
	ID	OOD	ID	OOD	ID	OOD	ID	OOD
Ours w/o Augmentations	46.4	31.2	92.3	91.4	73.1	30.1	93.6	54.3
Pixel-only	45.4	29.1	87.5	69.8	75.2	32.4	95.5	68.2
Frequency-only	**52.0**	35.6	98.3	95.7	73.4	37.6	95.1	77.9
Mask-low	51.2	36.0	98.3	96.1	**75.5**	38.4	95.1	80.5
Mask-high	51.5	35.4	97.9	95.4	75.1	36.4	95.0	79.5
Mask-ring	51.9	34.3	98.2	94.0	74.9	36.5	**95.7**	77.1
Ours	51.9	**36.8**	**98.6**	**96.4**	75.4	**40.7**	95.3	**83.4**

results show that the proposed dual-space targeted augmentation is dataset-agnostic and robust to different domain variations, while our focus remains on real-world deployment or environment shifts. The results with standard deviation are listed in Appendix Sec. A.

4.3 Ablation Study

Table 4 presents ablation results of the effects of pixel-space and frequency-space mixing and the impact of the gradient-guided mechanism. Across the four real-world datasets, our full method achieves the highest OOD performance, outperforming Ours w/o Augmentations and all single-modality variants. We also replace the gradient-guided map with fixed unguided maps with frequency masks: low-frequency mask, high-frequency mask, and ring mask. Frequency-only mixing consistently outperforms pixel-only mixing, highlighting the importance of frequency-space perturbation. Replacing the gradient-guided mechanism with maps with low/high/ring masks leads to a consistent drop in OOD performance, especially on BirdCalls and Galaxy10, confirming that this adaptive, sensitivity-aware guidance is essential for robust domain adaptation.

4.4 Generalization on Backbone Networks

To verify that D-GAP generalizes beyond backbone networks, we additionally evaluate with ConvNeXt-Tiny [32] and ViT-B/16 [8]. Table 5 shows that D-GAP consistently improves OOD performance on PACS, Digits-DG, and OfficeHome under both backbones. Fig. 6 further visualizes the ID–OOD trade-off on Galaxy10, where D-GAP remains SOTA performance on OOD robustness.

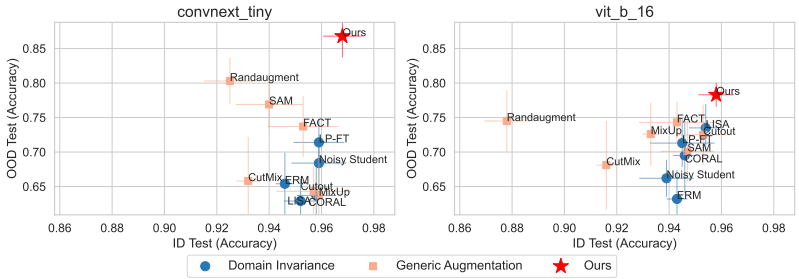


Figure 6: Performance of ConvNeXt and ViT on Galaxy10.

Table 5: Average OOD performance on PACS, Digits-DG, and OfficeHome with ConvNeXt and ViT.

Method	PACS		Digits		OfficeHome	
	ConvNeXt	ViT	ConvNeXt	ViT	ConvNeXt	ViT
SAM (2023) [65]	85.3	85.4	88.0	85.1	73.7	63.8
Su et al. (2024) [52]	87.9	75.3	88.9	85.6	73.9	64.6
AGST (2025) [62]	87.4	79.0	88.5	85.8	73.7	64.4
Ours	**88.7**	**86.4**	**89.8**	**86.8**	**75.0**	**66.0**

4.5 Domain Dependence of High-Gradient Frequency Locations

To further compare the overlap of high-gradient amplitude regions between same-class cross-domain (SC-CD) samples and different-class same-domain (DC-SD) samples, we compute the Jaccard similarity between the top-ranked high-gradient frequency locations. As shown in Table 6, DC-SD pairs consistently exhibit higher similarity across both datasets and all thresholds, suggesting that high-gradient frequency locations contain domain-dependent information rather than being determined solely by class identity. This observation motivates our gradient-guided amplitude mixing method.

4.6 Empirical Evaluations of Connectivity

To better understand why our method is effective on improving OOD robustness, we empirically evaluate the connectivity measures (details are in Appendix Sec. F). We present connectivity ratios α/γ and β/γ on datasets Camelyon17 and iWildCam in Table 7. Our method achieves the highest

Table 6: Jaccard similarity of top-ranked high-gradient frequency regions. SC-CD denotes same-class cross-domain pairs, while DC-SD denotes different-class same-domain pairs.

	Galaxy10		OfficeHome	
High-gradient region	SC-CD	DC-SD	SC-CD	DC-SD
Top 10%	0.5449	**0.5551**	0.2774	**0.3035**
Top 20%	0.5485	**0.5734**	0.3332	**0.3554**
Top 30%	0.5497	**0.5847**	0.4372	**0.4558**

α/γ value on both datasets, showing more consistent semantic alignment across domains as well as more effective randomization of spurious domain-dependent features $x_{d:\text{spu}}$. However, a moderate β/γ indicates our method may also slightly perturbs class-relevant features $x_{\text{obj}}, x_{d:\text{robust}}$. However, from Table 4 and Fig.5, our method still achieves the best OOD performance. The result is consistent with the estimation proposed by Shen et al. [49], where the target accuracy is estimated to be $(\alpha/\gamma)^{w_1} \cdot (\beta/\gamma)^{w_2}$, with higher weight on α/γ ($w_1 > w_2$). This observation further motivates our gradient-guided amplitude mixing method.

Table 7: Connectivity values across class, domain, and both on Camelyon17 and iWildCam datasets.

Camelyon17						**iWildCam**					
Augmentations	α	β	γ	α/γ	β/γ	**Augmentations**	α	β	γ	α/γ	β/γ
w/o Augmentations	0.015	0.189	0.002	7.50	94.5	w/o Augmentations	0.007	0.008	0.021	0.33	0.38
RandAugment	0.181	0.186	0.114	1.59	1.63	RandAugment	0.224	0.125	0.130	1.72	0.96
Stain Color Jitter	0.029	0.116	0.004	7.25	29.0	Copy-Paste	0.254	0.031	0.063	4.03	0.49
Ours	0.120	0.073	0.003	40.0	24.3	**Ours**	0.212	0.030	0.051	4.16	0.59

4.7 Computational Cost

Table 8 compares the computational performance and cost in a unified experimental setting. The additional time of D-GAP mainly comes from computing task gradients with respect to the Fourier amplitude in order to construct the model-response-adaptive sensitivity map.

Table 8: Computational cost comparison under the same setting.

Method	Time / Iter. (ms)	Throughput (img/s)	Peak GPU Mem. (GB)
ERM	24.46 ± 0.17	1308.3 ± 9.1	0.86
MixUp	24.71 ± 0.16	1294.9 ± 8.1	0.88
DANN	48.62 ± 0.24	658.2 ± 3.2	1.55
LISA	66.28 ± 0.19	482.8 ± 1.3	0.88
SAM	48.00 ± 0.30	666.6 ± 4.2	0.91
FACT	63.00 ± 0.31	507.9 ± 2.5	1.94
D-GAP (Ours)	90.46 ± 2.38	353.9 ± 9.2	0.86

5 Conclusion, Limitations and Future Work

We propose D-GAP, providing a simple and adaptive approach for OOD robustness in visual domain adaptation without dataset-specific analysis. By operating in both frequency and pixel spaces and adaptively interpolating amplitude values, our method achieves notable improvements on multiple real-world datasets and domain adaptation benchmark datasets. One limitation is that the current framework needs additional gradient computation. However, this cost comes from the model-response-adaptive nature of the method, which is also the key distinction from fixed or random Fourier mixing. Future work may reduce this overhead by updating sensitivity maps periodically, using low-resolution spectral proxies, or learning lightweight sensitivity predictors.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (NSFC) No. 62372393 and National Science and Technology Major Project of China (No. 2026ZD161120).

References

[1] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

[2] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021.

[3] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[6] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[7] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

[11] Irena Gao, Shiori Sagawa, Pang Wei Koh, Tatsunori Hashimoto, and Percy Liang. Out-of-domain robustness via targeted augmentations. In *International Conference on Machine Learning*, pages 10800–10834. PMLR, 2023.

[12] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

[13] Bruce C Hansen and Robert F Hess. Structural sparseness and spatial phase alignment in natural scenes. *Journal of the Optical Society of America A*, 24(7):1873–1885, 2007.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Xilin He, Jingyu Hu, Qinliang Lin, Cheng Luo, Weicheng Xie, Siyang Song, Muhammad Haris Khan, and Linlin Shen. Towards combating frequency simplicity-biased learning for domain generalization. *Advances in Neural Information Processing Systems*, 37:31078–31102, 2024.

[16] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021.

[17] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020.

[18] Leung Henry. Galaxy10 decals dataset., 2021.

[19] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018.

[20] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018.

[21] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[22] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6891–6902, 2021.

[23] Maximilian Ilse, Jakub M Tomczak, and Patrick Forré. Selecting data augmentation for simulating interventions. In *International conference on machine learning*, pages 4555–4562. PMLR, 2021.

[24] Alexis Joly, Hervé Goëau, Stefan Kahl, Lukáš Picek, Titouan Lorieul, Elijah Cole, Benjamin Deneu, Maximilien Servajean, Andrew Durso, Hervé Glotin, et al. Overview of lifeclef 2022: an evaluation of machine-learning based species identification and species distribution prediction. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 257–285. Springer, 2022.

[25] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4893–4902, 2019.

[26] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[28] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

[29] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

[30] Ingyun Lee, Wooju Lee, and Hyun Myung. Domain generalization with vital phase augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 2892–2900, 2024.

[31] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[32] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[33] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

[34] Ju-Hyeon Nam and Sang-Chul Lee. Fsda: Frequency re-scaling in data augmentation for corruption-robust image classification. *Pattern Recognition*, 150:110332, 2024.

[35] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 7. Granada, 2011.

[36] Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. Ssmba: Self-supervised manifold based data augmentation for improving out-of-domain robustness. *arXiv preprint arXiv:2009.10195*, 2020.

[37] A Oppenheim, Jae Lim, Gary Kopec, and SC Pohlig. Phase in speech and pictures. In *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 632–637. IEEE, 1979.

[38] Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981.

[39] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

[40] Hannah Pinson, Joeri Lenaerts, and Vincent Ginis. Linear cnns discover the statistical structure of the dataset using only the most dominant frequencies. In *International Conference on Machine Learning*, pages 27876–27906. PMLR, 2023.

[41] Leon N Piotrowski and Fergus W Campbell. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, 11(3):337–346, 1982.

[42] Helen Qu and Sang Michael Xie. Connect later: improving fine-tuning for robustness with targeted augmentations. In *Proceedings of the 41st International Conference on Machine Learning*, pages 41769–41786, 2024.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[44] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. *Advances in neural information processing systems*, 34:29935–29948, 2021.

[45] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, et al. Extending the wilds benchmark for unsupervised adaptation. In *International Conference on Learning Representations*, 2022.

[46] Tim Sainburg and Asaf Zorea. Noisereduce: Domain general noise reduction for time series signals. *arXiv preprint arXiv:2412.17851*, 2024.

[47] Sina Salehnia, Oznur Tastan, and Erchan Aptoula. Frequency domain image augmentation for domain generalized image classification. In *2025 33rd Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE, 2025.

[48] Ankita Sarkar, Sarbani Palit, and Ujjwal Bhattacharya. Demystifying galaxy classification: An elegant and powerful hybrid approach. In *2024 39th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE, 2024.

[49] Kendrick Shen, Robbie M Jones, Ananya Kumar, Sang Michael Xie, Jeff Z HaoChen, Tengyu Ma, and Percy Liang. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In *International conference on machine learning*, pages 19847–19878. PMLR, 2022.

[50] Patrice Y Simard, David Steinkraus, John C Platt, et al. Best practices for convolutional neural networks applied to visual document analysis. In *Icdar*, volume 3. Edinburgh, 2003.

[51] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[52] Jianjun Su, Xuejiao Yu, Xiru Wang, Zhijin Wang, and Guoqing Chao. Enhanced transfer learning with data augmentation. *Engineering Applications of Artificial Intelligence*, 129:107602, 2024.

[53] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. *Domain adaptation in computer vision applications*, pages 153–171, 2017.

[54] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer vision–ECCV 2016 workshops: Amsterdam, the Netherlands, October 8-10 and 15-16, 2016, proceedings, part III 14*, pages 443–450. Springer, 2016.

[55] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[56] David Tellez, Maschenka Balkenhol, Irene Otte-Höller, Rob Van De Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, et al. Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE transactions on medical imaging*, 37(9):2126–2136, 2018.

[57] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[58] Guangze Wang. Galaxy morphology classification with densenet. In *Journal of Physics: Conference Series*, volume 2580, page 012064. IOP Publishing, 2023.

[59] Ruoqi Wang, Haitao Wang, and Qiong Luo. Galaxalign: Mimicking citizen scientists' multimodal guidance for galaxy morphology analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 8636–8644, 2025.

[60] Shunxin Wang, Raymond Veldhuis, Christoph Brune, and Nicola Strisciuglio. What do neural networks learn in image classification? a frequency shortcut perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1433–1442, 2023.

[61] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dj Dvijotham, and Ali Taylan Cemgil. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*, 2021.

[62] Wai Keung Wong, Yuwu Lu, Junyi Wu, Zhihui Lai, and Xuelong Li. Approximate geometric structure transfer for cross-domain image classification. *Pattern Recognition*, 159:111105, 2025.

[63] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

[64] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[65] Chengchao Xu and Xinmei Tian. Semantic-aware mixup for domain generalization. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2023.

[66] Qinwei Xu, Ruipeng Zhang, Ziqing Fan, Yanfeng Wang, Yi-Yan Wu, and Ya Zhang. Fourier-based augmentation with applications to domain generalization. *Pattern Recognition*, 139:109474, 2023.

[67] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14383–14392, 2021.

[68] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020.

[69] Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, and Lei Zhang. Interactive self-training with mean teachers for semi-supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5941–5950, 2021.

[70] Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9011–9020, 2020.

[71] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4085–4095, 2020.

[72] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pages 25407–25437. PMLR, 2022.

[73] Mehmet Kerim Yucel, Ramazan Gokberk Cinbis, and Pinar Duygulu. Hybridaugment++: Unified frequency spectra perturbations for model robustness. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5718–5728, 2023.

[74] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[75] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[76] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13025–13032, 2020.

[77] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, pages 561–578. Springer, 2020.

Appendix

A Additional Comparison Results

Due to page limitations, we present the results in Tables 2 and 3 with the additional variance value over five random seeds here in Tables 11, 9, and 10. Results of DeepAll, FACT and SAM in these tables are copied from the source paper [65] for fair comparison, as our re-implementations under identical settings produced slightly lower numbers. As a result, we did not provide standard deviation values for DeepAll, FACT and SAM here for fair comparison.

Table 9: Model accuracy of leave-one-domain-out evaluation on Digits-DG (mean and standard deviation).

Methods	**MNIST**	**MNIST-M**	**SVHN**	**SYN**	**Avg.**
DeepAll	95.8	58.8	61.7	78.6	73.7
FACT (2023) [66]	97.9	65.6	72.4	90.3	81.5
SAM (2023) [65]	97.8	66.8	73.2	90.7	82.1
Su et al. (2024) [52]	97.3 ± 0.33	65.9 ± 0.51	74.5 ± 0.48	92.5 ± 0.31	82.6
AGST (2025) [62]	97.3 ± 0.14	63.1 ± 0.45	70.2 ± 0.78	93.2 ± 0.37	81.0
Ours	**98.3 ± 0.12**	**68.5 ± 0.26**	**77.2 ± 0.53**	**93.8 ± 0.24**	**84.5**

Table 10: Model accuracy of leave-one-domain-out evaluation on OfficeHome (mean and std).

Methods	Art	Clipart	Product	Real	Avg.
DeepAll	57.88	52.72	73.50	74.80	64.72
FACT (2023) [66]	60.34	54.85	74.48	76.55	66.56
SAM (2023) [65]	60.79	**55.47**	74.37	76.37	66.75
Su et al. (2024) [52]	64.13 ± 0.50	54.83 ± 0.49	75.60 ± 0.61	75.36 ± 1.04	67.71
AGST (2025) [62]	62.50 ± 0.61	53.87 ± 0.29	74.57 ± 0.48	75.70 ± 0.78	66.66
Ours	**64.82 ± 0.55**	55.21 ± 0.37	**77.76 ± 0.43**	**83.08 ± 0.25**	**70.22**

Table 11: Model accuracy of leave-one-domain-out evaluation on PACS (mean and std).

Methods	**Art**	**Cartoon**	**Photo**	**Sketch**	**Avg.**
DeepAll	84.94	76.98	97.64	76.75	84.08
FACT (2023) [66]	**89.68**	81.06	97.02	83.76	87.88
SAM (2023) [65]	85.25	82.02	96.75	85.04	87.27
Su et al. (2024) [52]	79.88 ± 0.53	78.03 ± 0.57	95.93 ± 0.21	73.15 ± 0.39	81.75
AGST (2025) [62]	83.50 ± 0.76	80.83 ± 0.48	96.49 ± 0.23	79.57 ± 0.62	85.35
Ours	89.34 ± 0.34	**82.81 ± 0.63**	**97.99 ± 0.14**	**85.98 ± 0.19**	**89.03**

B Analysis over Hyperparameters

Effect of the frequency-window ratio. We further investigate the influence of the frequency-window side-length ratio r, which controls the size of the square region Ω_r used for amplitude mixing in the Fourier space (in Equation 3). As shown in Fig. 7, the choice of r has only a minor effect on the in-domain (ID) performance. On both Galaxy10 and BirdCall, the ID accuracy/F1 remains relatively stable across different values of r. For out-of-domain (OOD) generalization, the performance is more sensitive to r, but the variation remains moderate as long as the frequency window is not zero. In particular, both datasets achieve strong OOD performance when r is in the range of 0.4 to 0.6. Although other none-zero choices of r are slightly inferior to this range, they still consistently outperform competing baseline methods, demonstrating that the proposed augmentation is robust to the exact selection of the window size. In the experiments in our paper, we uniformly set $r = 0.5$ for all datasets, with which our method achieves state-of-the-art results. And we recommend using $r = 0.5$ as the default setting in further practical applications.

Effect of the mixing ratio λ_1 and λ_2. We analyze the sensitivity of the proposed method to the mixing ratios λ_1 and λ_2 by varying both values from 0 to 1. As shown in Fig. 8, the ID performance remains stable across different parameter pairs, suggesting that the proposed augmentation does

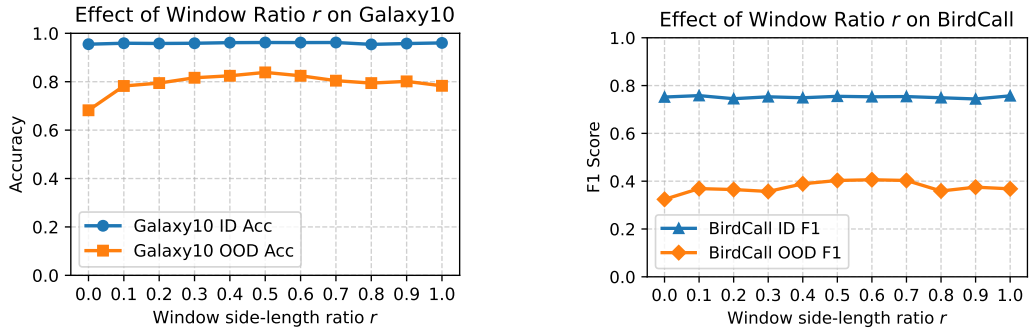


Figure 7: Effect of the frequency-window side-length ratio r on Galaxy10 and BirdCalls datasets.

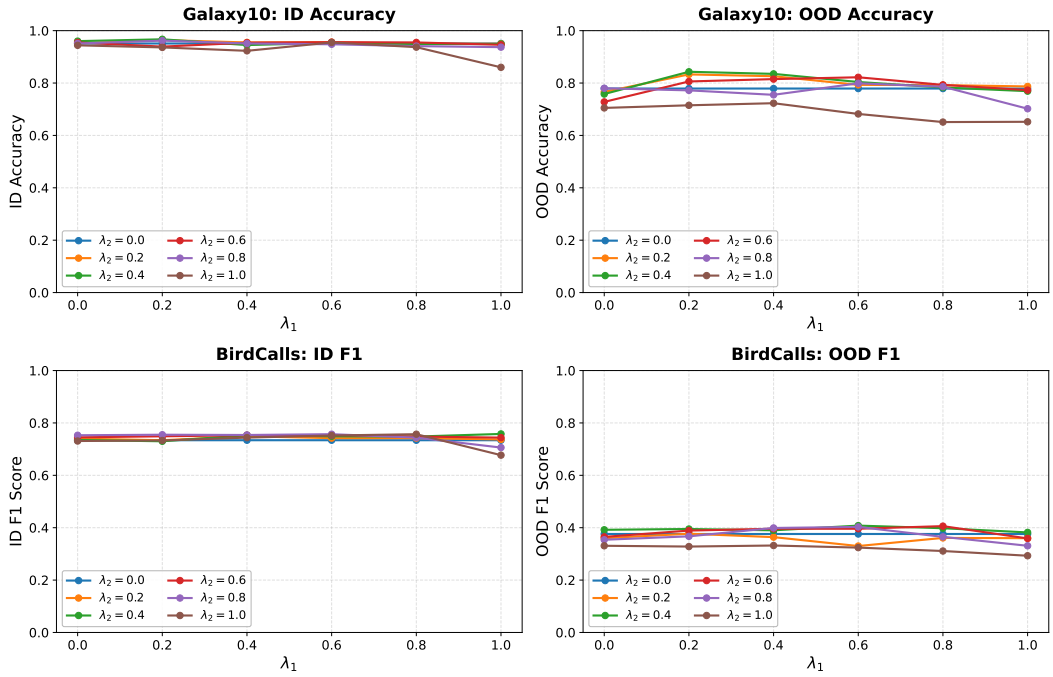


Figure 8: Effect of the mixing ratios λ_1 and λ_2 on Galaxy10 and BirdCalls datasets.

not degrade source-domain learning. The OOD performance is slightly more sensitive, but remains consistently strong as long as λ_1 and λ_2 are not set to extreme values such as 0 or 1. Most non-extreme combinations achieve competitive results and still outperform the baseline methods, indicating that the method is robust to the exact choice of these two hyperparameters. Overall, we recommend setting $\lambda_1 \in [0.2, 0.6]$ and $\lambda_2 \in [0.2, 0.6]$, while avoiding overly weak or overly aggressive mixing.

C Related Work

C.1 Pretraining for OOD Robustness

Pretraining on large-scale unlabeled data has emerged as a promising OOD generalization method [4, 49, 43, 45]. Contrastive pretraining, which encourages feature alignment across domains using generic augmentations, outperforms traditional domain adaptation methods that enforce explicit domain invariance [25, 19]. However, the standard fine-tuning with generic augmentations often leads to unsatisfying OOD performance due to misalignment with specific distribution shifts [42, 28]. Although Connect Later [42] improves OOD robustness by applying targeted augmentations during

linear probing then fine-tuning (LP-FT) [28], it still relies heavily on manually designed augmentations and cannot easily generalize without shift-specific analysis.

C.2 Augmentations

For a given input $x \in \mathcal{X}$, augmented variants $x' \sim \mathcal{A}(\cdot \mid x)$ are often used to enhance robustness [17, 16]. Data augmentation has been widely used to improve model robustness to label-invariant transformations such as translation or rotation [17, 44, 36]. Traditional augmentation strategies typically apply generic perturbations to increase input diversity [50, 27, 6, 7, 75]. However, recent studies highlight that the choice of augmentation type significantly impacts model generalization [5, 63].

In semi-supervised learning, augmentations have also been utilized in the self-training paradigm to improve generalization by training on pseudo-labeled data [64, 51, 69]. Nevertheless, Qu et al. [42] show that a self-training baseline using pseudo-labels generated from an ERM model with targeted augmentations still underperforms compared to Connect Later [42]. This underscores the importance of pretrained representations in achieving robustness gains. Connect Later [42] leverages targeted augmentations not only as a data transformation tool, but as a design interface to incorporate knowledge of the distribution shift while still benefiting from pretrained features.

Recent work in domain shift problems emphasizes the effectiveness of *targeted augmentations*—those informed by the nature of the distribution shift—over generic or even target-aware augmentations such as CutMix [11, 74]. While [11] focuses on domain generalization where the target domain is unknown, some recent work [42, 11] studies the domain adaptation setting with access to unlabeled target data. The targeted augmentations are effective but rely on manually designing and are dataset-specialized.

Frequency-domain augmentation has been widely used to improve robustness under domain shifts. FDA [71] replaces the low-frequency amplitude of source images with target-domain spectra, and [70] preserves phase for semantic consistency, but both rely on manually constrained spectral rules. FSDR [22] randomizes domain-variant frequency components, but its perturbation is based on domain-level statistics rather than the model's local task sensitivity. HybridAugment++ [73] unifies multiple spectrum perturbations, FSDA [34] learns frequency re-scaling from spectral patterns, and Vital Phase Augmentation [30] augments phase variations; however, these methods do not explicitly inject target-domain information according to frequency-wise model responses. FACT [66] mixes Fourier amplitudes to encourage phase-preserved semantic learning, while SAM [65] further introduces semantic-aware amplitude/phase mixup; however, their mixing strategies are still predefined rather than model-response-adaptive. In conclusion, existing methods [65, 66, 71, 70, 73, 34, 30, 47, 22] typically use fixed, random, manually constrained, or data-statistic-driven perturbations, limiting targeted domain intervention. In contrast, D-GAP computes task-gradient-based amplitude sensitivity maps to adaptively control frequency-wise target-domain interpolation, and further combines pixel-space blending to compensate local spatial details, enabling more targeted and flexible OOD augmentation.

C.3 On the Importance of Phase Information

A series of early works [37, 38, 41, 13] have highlighted that in the Fourier spectrum of visual signals, the phase component is critical for preserving high-level semantics, while the amplitude primarily captures low-level appearance statistics. Recent advances further explore this perspective in domain adaptation. For example, [71] proposes to replace a local region in the centralized amplitude spectrum of a source image with that of a target image to synthesize target-style images for training. In parallel, [70] introduces phase consistency as a constraint during source-target translation, demonstrating its superiority over traditional cycle consistency [20] in domain adaptation tasks, especially for semantic segmentation. Xu et al. [66] incorporate a Fourier-based strategy that emphasizes phase preservation to enhance generalization performance under domain shift. Inspired by this, we propose to use phase in frequency space to preserve the task-relevant features in our augmentations.

D Dataset Details

Species Classification from Camera Trap Images (iWildCam). In the iWildCam dataset [2, 26], the task is to classify an animal species y from images x captured by static camera traps. The

dataset consists of 243 camera locations in D^{train}. Although images from the same camera share similar backgrounds, domain-dependent features $x_{d:robust}$, such as habitat, vary across domains. For example, cameras in Kenyan savannas (cameras 23 and 97) and a Guatemalan forest (camera 54) have differing background features. This requires the model to focus on foreground features (e.g., animal species) for accurate prediction, as background habitat often doesn't provide useful information for classification [11, 42].

Tumor Identification in Histopathology Slides (Camelyon17). In Camelyon17 [2, 26], the task is to classify tumor presence in histopathology slides. These slides come from three hospitals, each with varying imaging protocols and staining techniques that lead to domain-specific color variations. For example, some hospitals have mostly early-stage pN1 breast cancer, while others have later-stage pN2 cases. These domain-specific features, such as tumor size and density, correlate with the cancer stage y, requiring the model to account for these domain shifts in order to generalize effectively [11, 42].

Bird Species Recognition from Audio Recordings (BirdCalls). The BirdCalls dataset [24, 11] consists of audio recordings from nine microphones located in different regions. The task is to classify the bird species y vocalizing in each audio clip x. While the bird species information is crucial, low-level noise and microphone settings (e.g., gain levels) often introduce irrelevant domain-specific features. These features, such as background noises from different habitats, should be ignored by the model, which should focus on bird calls for accurate classification. Randomizing microphone-related features ensures that the model generalizes across diverse recording environments [11].

Astronomical Morphology Classification (Galaxy10 DECaLS&SDSS). Galaxy 10 DECaLS and SDSS are two galaxy morphology classification datasets containing images from two different surveys [18]. They differ in image quality, resolution, and color distribution due to differences in telescope instrumentation. We treat Galaxy10-DECaLS as the source domain and Galaxy10-SDSS as the target domain, and evaluate cross-domain generalization without using any target-domain labels during training.

Additional Benchmark Datasets. In addition to the four real-world datasets used throughout this paper, we further evaluate our method on three commonly used domain adaptation/generalization benchmarks: **PACS** [31], **Office-Home** [57], and **Digits-DG** [76]. Compared to our four *real-world* datasets that cover diverse modalities and deployment-related shifts (wildlife, histopathology, audio, and astronomical imaging), these common benchmarks focus on visual recognition with well-studied style variations (PACS: {Art, Cartoon, Photo, Sketch}; Office-Home: {Art, Clipart, Product, Real-World}; Digits-DG: multiple digits domains including MNIST [29], MNIST-M [9], SVHN [35], and SYN [9]).

E Experimental Details

E.1 D-GAP

We listed some important implementation details of our method as follows. Other implementation details can be find in our code in the supplementary material. And we will release the code after this paper is published.

iWildCam. We use ResNet-50 [14] pretrained with SwAV [4] as the backbone [42] in Section 4.2 and 4.3. For pretrained models, we follow LP-FT [28, 42]: a linear probe is trained for 10 epochs using Adam, and its weights are used to initialize the full model, which is then fine-tuned for another 20 epochs. Following [42], we perform 10 hyperparameter search trials by sampling: linear probe learning rate $10^{\text{Uniform}[-3,-2]}$, fine-tuning learning rate $10^{\text{Uniform}[-5,-2]}$, and augmentation probability from Uniform$[0.5, 0.9]$. The configuration with the best OOD validation performance is selected for testing. To mitigate outlier variance, our method is averaged over 5 seeds. Other hyperparameters are the same as those of Qu et al. [42].

Camelyon17. We conduct experiments on the unlabeled Camelyon17 dataset [45] using DenseNet121 [21] pretrained with SwAV [4] in Sections 4.2 and 4.3, following the protocol in [42, 11]. For pretrained models, we first apply 10 epochs of linear probing before full fine-tuning

Table 12: Hyperparameter settings for baseline methods on Galaxy10.

Method	Hyperparameters
ERM	Learning rate $= 10^{-4}$ Weight decay $= 0$
LISA	Learning rate $= 10^{-4}$ Weight decay $= 0$ Transform probability $= 0.9$
MixUp	Learning rate $= 10^{-4}$ Weight decay $= 0$ Transform probability $= 0.9$ $\alpha = 0.4$
CutMix	Learning rate $= 10^{-4}$ Weight decay $= 0$ Transform probability $= 0.9$ $\alpha = 0.4$
Cutout	Learning rate $= 10^{-4}$ Weight decay $= 0$ Transform probability $= 0.9$ $\alpha = 0.4$
Randaugment	Learning rate $= 10^{-4}$ Transform probability $= 0.9$ Version = $Original$
CORAL	Learning rate $= 10^{-4}$ Coral weight $= 1.0$
Noisy Student	Learning rate $= 10^{-4}$ soft label $= False$
DANN	Classifier learning rate $= 10^{-4}$ Discriminator learning rate $= 10^{-4}$ GRL$\lambda = 2/(1 + exp(-10epoch/15)) - 1$

for 20 epochs, and select the linear probe checkpoint with the highest OOD validation accuracy for initialization. For hyperparameter tuning, following [42], the following hyperparameters are sampled: linear probe learning rate from $10^{\text{Uniform}[-3,-2]}$, fine-tuning learning rate from $10^{\text{Uniform}[-5,-2]}$, augmentation probability from Uniform$[0.5, 0.9]$, and augmentation strength from Uniform$[0.05, 0.1]$. The best hyperparameter configuration is selected based on OOD validation performance. All results are averaged over 5 random seeds. Other hyperparameters are the same as those of Qu et al. [42].

BirdCalls. We evaluate our method on the BirdCalls dataset using a pretrained EfficientNet-B0 [55] encoder in Sections 4.2 and 4.3. Following [11], we first perform 10 epochs of linear probing (LP-FT [28]) with the Adam optimizer, selecting the checkpoint with the best OOD validation performance to initialize the fine-tuning stage for 30 epochs. Consistent with the protocol in [42], we sample the linear probe learning rate from $10^{\text{Uniform}[-3,-2]}$, the fine-tuning learning rate from $10^{\text{Uniform}[-4,-3]}$, the augmentation probability from Uniform$[0.5, 0.9]$. The final hyperparameter setting is selected based on the best OOD validation performance. Reported results are averaged over 5 random seeds. Other hyperparameters are the same as those of Gao et al. [11].

Galaxy10. We evaluate our method on the Galaxy10 DECaLS and Galaxy10 SDSS datasets with the ResNet18 [14] backbone network in Sections 4.2 and 4.3. Specifically, we use the DECaLS for training and validation, and evaluate the model on the SDSS subset, where labels are mapped to DECaLS categories using a predefined label conversion rule since the labels in the two datasets are not exactly the same. Our training pipeline employs a ResNet18 encoder initialized from a contrastively pre-trained checkpoint, followed by a linear classifier trained end-to-end. We first perform 10 epochs of linear probing (LP-FT [28]) with the Adam optimizer, selecting the checkpoint with the best OOD validation performance to initialize the fine-tuning stage. The fine-tuning stage lasts for 15 epochs using the Adam optimizer with a learning rate of 10^{-4} and a batch size of 8. The augmentation probability is set to 0.5.

Benchmark Datasets. And in Sections 4.2 and 4.3, we use the same ConvNet backbone as [76, 65, 77, 66] for Digits-DG, and ResNet18 [14] for OfficeHome and PACS. Please find the details in the original paper and the released source code.

In D-GAP, the target image for mixing can be selected from either labeled or unlabeled data. For consistency and fair comparison with baseline methods [11, 42], we adopt a unified strategy across datasets. Specifically, for iWildCam, Camelyon17, and BirdCalls, we mix each source image with samples from other domains within the training set. For Galaxy10, since the training data only contains images from a single telescope, we randomly sample images from the unlabeled test set as mixing targets to generate cross-domain augmented images. For Digits-DG, OfficeHome, and PACS, we follow the target-source splitting as in [65, 66]. And the unlabeled target image for mixing can be selected from the held-out target domain during training.

E.2 Baseline Methods

For baseline methods on iWildCam, Camelyon17, BirdCalls dataset, we followed the settings and hyperparameters exactly in [42, 11]. For the baseline methods for Digits-DG, OfficeHome, and PACS, we follow the settings in their original papers [65, 66, 62, 52]. And for Galaxy10 dataset, the settings are presented in Table 12.

F Computation of Connectivity

To compute the average connectivity between two class-domain pairs (c, d) and (c', d'), the process involves labeling all training examples from class c and domain d as 0, and those from class c' and domain d' as 1, while discarding examples from other classes or domains. Then, a ResNet50 model is trained using augmentations, stochastic gradient descent (SGD) with momentum, and a cosine learning rate schedule, maintaining consistent training parameters across all classes and domains. Specifically, we train binary classifiers from scratch to predict the class-domain pair of each input example. The test set is prepared similarly to the training set, and the model's performance on the test dataset provides an estimate of connectivity between the specified class-domain pairs, quantified by the test error rate. For iWildCam, we randomly select 15 class-domain pairs, while for Camelyon17, we use all class-domain pairs since it is a binary classification task. We train using the same hyperparameters described in E for 3,000 steps, applying early stopping based on validation accuracy. We will release the code of connectivity computing after the paper is published.

G Broader Impact

This work studies dataset-agnostic augmentation for improving out-of-domain robustness in visual domain adaptation. Its potential positive impact lies in improving the reliability of machine learning models under distribution shifts, especially in scientific, ecological, and medical imaging scenarios where collecting labeled target-domain data is costly. By reducing the need for dataset-specific expert augmentation design, the proposed method may lower the barrier for deploying robust models in diverse real-world environments.

Potential negative impacts are limited but should not be ignored. Since the method improves model robustness under domain shifts, it could also be applied to sensitive visual recognition systems. In high-stakes domains such as medical imaging, improved OOD performance does not guarantee clinical reliability, and models should not be deployed without domain-specific validation, uncertainty estimation, and human expert oversight.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly states the claims made, including the contributions made in the paper and important assumptions and limitations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Discussed in Section 5 Conclusion, Limitations and Future Work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: It is an empirical paper that proposes a new method and justifies it through experimental results, intuitive explanations, and references to existing theoretical frameworks (e.g., feature decomposition and connectivity).

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Experimental setups are reported in section Experiments and the Appendix. The code is provided in the supplementary material. And we will release the code after the paper is accepted.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code and the access to the datasets is provided in the supplementary material. All the code we used for the experiments will be public if the paper is accepted.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Some details are in Section Experiments. Other details are provided in our appendix and code in the supplemental material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide error bars in our experimental results and variance variance over multiple seeds in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: This paper provides the compute resources in Section 4.1-Experimental Settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss potential positive impacts in improving robust visual recognition for scientific, ecological, and medical imaging, as well as potential risks when the method is used in sensitive visual recognition or high-stakes decision systems (Appendix Sec. G).

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The original paper that produced the code package or dataset are cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: See supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.