PhaseAT: Fourier Phase Adversarial Training for Medical Image Domain Generalization

Ahmed Sharshar (✉), Asif Hanif, Naveen Kumar Kummari, Mohammad Yaqub, and Mohsen Guizani

Division of Computing and Mathematical Sciences
Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
{ahmed.sharshar, asif.hanif, naveen.kummari, mohammad.yaqub, mohsen.guizani}@mbzuai.ac.ae

Abstract. Reliable clinical deployment of deep medical image models is hindered by distribution shifts across scanners, sites, and acquisition protocols. Existing domain generalization (DG) methods often focus on style or intensity diversification, but they can still leave networks dependent on domain-specific texture correlations. Inspired by evidence that Fourier phase encodes semantic structure, we introduce PhaseAT, a phase-aware adversarial training framework for medical DG. PhaseAT forms phase-perturbed training views in the Fourier domain by iteratively updating a bounded phase perturbation while keeping the amplitude spectrum unchanged, thereby stressing spatial organization under matched appearance statistics. Perturbations are applied only to the luminance channel in YCbCr color space to avoid chromatic artifacts. Additionally, a simple phase-saliency mask concentrates updates on the most influential frequencies. The model is trained with a weighted combination of losses on clean and phase-perturbed samples, supporting both single-source and multi-source DG. We validate our method on two challenging medical datasets and demonstrate that PhaseAT achieves over 20% improvement in single-source domain generalization, outperforming several state-of-the-art DG methods. The code implementation is available at: https://github.com/ahmed-sharshar/PhaseAT

Keywords: Domain Generalization · Adversarial Training · Phase-Aware Perturbation.

1 Introduction

Deep learning has become the dominant paradigm for medical image analysis, yet standard training pipelines still rely on an *i.i.d.* assumption that rarely holds in clinical practice. In reality, medical imaging systems routinely experience distribution shifts induced by scanner vendors, acquisition protocols, reconstruction pipelines, and diverse patient populations [17]. From a spectral viewpoint, such shifts are often reflected in changes to the Fourier *amplitude* spectrum, which captures low-level appearance statistics (e.g., intensity and texture), while the Fourier phase largely preserves spatial organization and geometry [13]. Since

standard CNNs are biased toward texture cues [8], they frequently overfit to these unstable amplitude-driven correlations, causing brittle performance when deployed outside the source environment.

Domain Generalization (DG) targets this challenge by learning predictors that remain accurate on unseen target domains without requiring target data during training. This distinction is crucial in medicine, where privacy constraints and logistical silos limit multi-institutional data aggregation, often leaving practitioners with only a small number of source environments [18]. Multi-center benchmarks further underscore that such shifts are routine rather than exceptional in medical imaging, with substantial gaps under cross-hospital evaluation [4]. The central DG question remains: how can we learn features that stay stable when imaging style changes but clinical semantics do not?

Prior medical DG methods have made steady progress, though most target only subsets of shift factors. Existing approaches broadly fall into three families. First, *style/intensity diversification* methods simulate scanner-driven appearance variation, for example via adversarial style composition [5], content–style augmentation [6], or monotonic adversarial intensity attacks [22]; others use representation-driven objectives to regularize style/content separation [14]. Second, *frequency-domain* approaches exploit spectral structure through Fourier mixing, amplitude synthesis, or reconstruction constraints to reduce scanner-specific artifacts [16]. Third, *structure-consistent* methods that focus on anatomy-aligned cues, such as via anatomy-consistent pseudo-modalities [10] or edge-guided generalization [11]. Despite these advances, performance can remain sensitive to hyperparameter choices, and carefully tuned baselines are often surprisingly strong [19]. This motivates mechanisms that more directly reflect how medical shifts arise in practice [23], especially in limited-source regimes where the inductive bias of the training objective becomes paramount [20].

A promising route to robustness is *Adversarial Training* (AT), but the perturbation model must reflect clinically relevant shifts. Standard pixel-space perturbations tend to emphasize high-frequency noise, potentially reinforcing a model's texture bias [8]. A Fourier viewpoint enables more targeted control: the *amplitude* spectrum is closely tied to low-level appearance statistics, while *phase* governs spatial organization [13]. Indeed, phase is the primary carrier of recognizable structure [24]. Recent work further shows that improving stability to phase manipulations can enhance structural robustness [24], and frequency-constrained perturbations can effectively expose model vulnerabilities in medical tasks [21]. Together, these observations suggest that a *phase-focused* robustness objective can better match clinical distribution shifts while discouraging shortcut learning.

In this work, we propose *Phase-Aware Adversarial Training* (PhaseAT) for medical domain generalization, Figure 1. PhaseAT synthesizes training views by adversarially perturbing the Fourier *phase*, thereby challenging anatomy-relevant spatial organization, while *fixing* the Fourier amplitude to preserve per-frequency energy statistics (appearance). To avoid unnecessary chromatic artifacts and target geometry directly, we apply perturbations only to the luminance channel in YCbCr. The resulting phase-perturbed samples serve as hard training instances

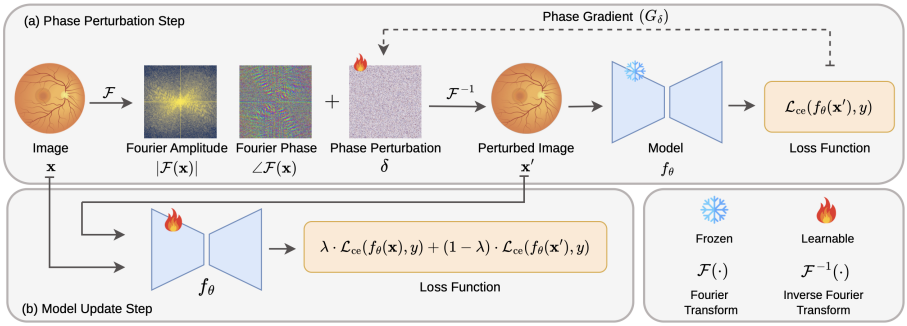

Fig. 1: PhaseAT Pipeline **(a)** *Phase Perturbation Step (inner maximization):* Adversarial phase perturbations are applied in the Fourier domain by updating the phase component of the luminance channel, generating phase-perturbed view x′. **(b)** *Model Update Step (outer minimization):* Model parameters are optimized using a joint objective that incorporates both clean and phase-perturbed samples to improve robustness.

that complement recent appearance diversification and Fourier-based DG strategies [22,16] by explicitly encouraging invariance to geometry-focused spectral variations under comparable appearance statistics. Unlike Fourier-based DG that varies the *amplitude* spectrum [16] or pixel-space augmentation, PhaseAT fixes amplitude and perturbs only *phase*, isolating geometry from appearance.

Contributions. We propose PhaseAT, a phase-aware adversarial training for medical image *single-source* and *multi-source* domain generalization, using robustness as a mechanism to improve out-of-domain performance. We introduce a *phase-constrained* spectral perturbation scheme that perturbs Fourier phase and applies perturbations in luminance to reduce artifacts under domain shift. Finally, we provide a comprehensive evaluation across different datasets, demonstrating consistent improvements over a range of recent DG baselines.

2 Methodology

To instantiate the phase-aware robustness objective described in Section 1, we generate phase-perturbed samples in the Fourier domain while preserving the amplitude spectrum, and incorporate these samples during training to enhance domain generalization. Specifically, spectral perturbations are applied only to the luminance channel in the YCbCr color space. This design choice is motivated by three considerations. First, YCbCr allocates higher effective resolution to luminance than to chrominance, reflecting the fact that fine spatial details are primarily encoded in the (Y) channel. Second, human visual sensitivity to high-frequency spatial structure is substantially greater for achromatic (luminance) variations than for chromatic ones, meaning that edges and geometric information are predominantly conveyed by luminance [15]. Third, prior studies have

shown that gradient-based perturbations tend to concentrate in the (Y) channel when images are represented in YCbCr, suggesting that restricting the inner maximization to luminance is both effective and avoids unnecessary chromatic distortions [15]. From an optimization perspective, operating on a single channel (Y) rather than three (RGB) reduces the dimensionality of the inner maximization problem. This not only improves computational efficiency but also produces more structured perturbations with fewer color artifacts.

2.1 Phase-Constrained Spectral Perturbations

Consider an RGB image $\mathbf{x} \in [0,1]^{H\times W\times 3}$ and its corresponding label y, with pixel values normalized to $[0,1]$ for each channel. The RGB image $\mathbf{x}$ is first converted into the YCbCr color space, producing $(\mathbf{x}_{\mathrm{Y}}, \mathbf{x}_{\mathrm{Cb}}, \mathbf{x}_{\mathrm{Cr}}) \in [0,1]^{H\times W\times 3}$, where the $\mathbf{x}_{\mathrm{Y}}$ is luminance channel. The luminance channel is transformed into the Fourier domain, yielding the complex spectrum $\mathrm{X} = \mathcal{F}(\mathbf{x}_{\mathrm{Y}}) \in \mathbb{C}^{H\times W}$, which can be decomposed into amplitude and phase components as $\mathrm{X} = |\mathrm{X}|e^{(j\angle \mathrm{X})}$. To generate a phase-perturbed representation, we introduce a perturbation $\delta \in \mathbb{R}^{H\times W}$ and define the modified spectrum as: $\mathrm{X}' = |\mathrm{X}|e^{j(\angle \mathrm{X}+\delta)}$, thereby preserving the original amplitude spectrum $|\mathrm{X}|$ while altering only the phase component. Applying the inverse Fourier transform yields the perturbed luminance channel, $\mathbf{x}'_{\mathrm{Y}} = \mathcal{F}^{-1}(\mathrm{X}')$. The perturbed luminance channel is combined with the original chrominance channels to form $(\mathbf{x}'_{\mathrm{Y}}; \mathbf{x}_{\mathrm{Cb}}; \mathbf{x}_{\mathrm{Cr}})$, which is then transformed back to RGB space, yielding the perturbed image $\mathbf{x}'$. To obtain adversarial phase perturbations, δ is optimized by maximizing the cross-entropy loss:

$$\underset{\delta}{\text{maximize}} \;\; \mathcal{L}_{\mathrm{ce}}(f_\theta(\mathbf{x}'), y) \quad \text{s.t.} \quad \|\delta\|_\infty \le \epsilon, \tag{1}$$

where ϵ specifies the maximum admissible phase deviation. Since X is complex-valued, the phase update is computed directly in the Fourier domain. Specifically, we derive the gradient of the loss with respect to the spectrum phase and use the resulting phase-direction gradient as a proxy signal for updating δ. Using Wirtinger calculus [12], phase-direction proxy gradient G_δ can be expressed as:

$$G_\delta \triangleq -2 \cdot \Im(G_{\mathrm{X}} \odot \mathrm{X}), \tag{2}$$

where G_{X} denotes the gradient of the loss w.r.t the complex spectrum X, $\odot$ is the Hadamard product and $\Im(\cdot)$ denotes the imaginary part. To ensure a stable and well-defined phase update, we compute a phase-saliency soft mask $W_\delta \in [0,1]^{H\times W}$ by applying min–max normalization to the magnitude of the phase-direction gradient, $|G_\delta|$. This mask captures the loss sensitivity to phase perturbations, upweighting frequency components that induce larger loss changes. It thus focuses the phase update on structurally informative, adversarially relevant frequencies. Phase perturbation is updated as follows:

$$\delta \;\leftarrow\; \mathcal{P}_{[-\epsilon,\epsilon]}\Big(\delta + \eta \cdot W_\delta \odot \mathrm{sign}(G_\delta)\Big), \tag{3}$$

where η denotes the update step size, $W_\delta \in [0,1]^{H\times W}$ is the phase-saliency weighting mask, $\mathcal{P}_{[-\epsilon,\epsilon]}(\cdot)$ denotes element-wise projection onto the ℓ_∞ ball of radius ϵ, and $\mathrm{sign}(\cdot)$ is the element-wise sign operator.

Algorithm 1 PhaseAT: Phase-Adv. Training for Domain Generalization

Source Data $\mathcal{D}_s$, Model f_θ, Epochs N, Steps K, Step Size η, Budget ϵ, Loss Weight λ, Learning Rate γ, FFT $\mathcal{F}$, IFFT $\mathcal{F}^{-1}$, Projection $\mathcal{P}_{[-\epsilon,\epsilon]}$, Imaginary Part $\Im$

for epoch $= 1, \ldots, N$ **do**
  $(\mathbf{x}, y) \sim \mathcal{D}_s$ ▷ Sample a mini-batch (single-source or pooled multi-source)
  $(\mathbf{x}_\mathrm{Y}, \mathbf{x}_\mathrm{Cb}, \mathbf{x}_\mathrm{Cr}) \leftarrow \texttt{RGB2YCbCr}(\mathbf{x})$ ▷ Convert from RGB to YCbCr color space
  $\mathrm{X} \leftarrow \mathcal{F}(\mathbf{x}_\mathrm{Y}) = |\mathrm{X}|e^{j\angle \mathrm{X}}$ ▷ Luminance channel spectrum
  $\delta \leftarrow \mathbf{0} \in \mathbb{R}^{H \times W}$ ▷ Initialize phase perturbation
  for $k = 1, \ldots, K$ **do**
    $\mathrm{X}' \leftarrow\!\shortmid |\mathrm{X}|e^{j(\angle \mathrm{X} + \delta)}$ ▷ Add perturbation to phase
    $\mathcal{F}^{-1}(\mathrm{X}') \rightarrow \mathbf{x}'_\mathrm{Y} \rightarrow \texttt{YCbCr2RGB}([\mathbf{x}'_\mathrm{Y}, \mathbf{x}_\mathrm{Cb}, \mathbf{x}_\mathrm{Cr}]) \rightarrow \mathbf{x}'$ ▷ Fourier-to-Spatial
    Compute cross-entropy loss: $\mathcal{L}_\mathrm{ce}(f_\theta(\mathbf{x}'), y)$
    $G_\delta \triangleq -2 \cdot \Im(\nabla_\mathrm{X} \odot \mathrm{X})$ ▷ Phase-direction gradient (Eq. 2)
    $\delta \leftarrow\!\shortmid \mathcal{P}_{[-\epsilon,\epsilon]}\big(\delta + \eta \cdot W_\delta \odot \mathrm{sign}(G_\delta)\big)$ ▷ Update δ with gradient ascent (Eq. 3)
  end for
  $\ell \leftarrow\!\shortmid \lambda\, \mathcal{L}_\mathrm{ce}(f_\theta(\mathbf{x}), y) + (1-\lambda)\, \mathcal{L}_\mathrm{ce}(f_\theta(\mathbf{x}'), y)$ ▷ Training loss (Eq. 4)
  $\theta \leftarrow\!\shortmid \theta - \gamma \nabla_\theta \ell$ ▷ Update model parameters via gradient descent
end for

Return: Adversarially trained robust model f_θ

2.2 Phase-Aware Adversarial Training

Given a clean sample $\mathbf{x}$ and its phase-perturbed counterpart $\mathbf{x}'$, we update model parameters by minimizing a convex combination of clean and perturbed losses:

$$\underset{\theta}{\text{minimize}} \quad \lambda \cdot \mathcal{L}_\mathrm{ce}(f_\theta(\mathbf{x}), y) + (1-\lambda) \cdot \mathcal{L}_\mathrm{ce}(f_\theta(\mathbf{x}'), y), \tag{4}$$

where $\lambda \in [0, 1]$ controls the trade-off between standard and robust accuracy. Algorithm 1 summarizes the full training procedure; the same routine applies to single-source and multi-source DG by defining $\mathcal{D}_s$ as the available source data.

3 Experimental Setup

Datasets. We evaluate on two medical benchmarks with strong cross-domain variation. *Camelyon17-WILDS* [4] contains histopathology data from *five medical centers* (domains). *Diabetic Retinopathy (DR)* combines Aptos [3], EyePACS [2], Messidor [7], and Messidor-2 [1] as distinct domains for multi-class severity prediction.

Evaluation Protocol. We follow the *DomainBed* methodology [9] which uses training-domain validation for model selection. We report accuracy on unseen domains in two regimes: (1) Single-domain generalization (SDG), where we train on one source and evaluate on the rest, and (2) Multi-source generalization (MSDG), using a leave-one-domain-out protocol.

Table 1: Performance comparison in the Single Domain Generalization (SDG) setting. Values are the accuracy(%) mean ± standard deviation. Best and second-best per column are in **bold** and underline; † marks a significant gain over the strongest baseline (paired t-test, $p<0.05$).

(a) Camelyon17-WILDS

Method	C_0	C_1	C_2	C_3	C_4	AVG.
Source Domain: C_0						
ERM	–	$67.29_{\pm 1.2}$	$65.89_{\pm 2.1}$	$67.73_{\pm 1.5}$	$56.50_{\pm 2.8}$	$64.35_{\pm 1.9}$
AdverIN [22]	–	$70.17_{\pm 1.4}$	$69.30_{\pm 2.3}$	$69.92_{\pm 1.7}$	$60.26_{\pm 2.5}$	$67.41_{\pm 2.0}$
ConDiSR [14]	–	$69.05_{\pm 1.1}$	$67.45_{\pm 2.0}$	$70.29_{\pm 1.6}$	$\underline{60.48}_{\pm 2.4}$	$66.82_{\pm 1.8}$
RAS^{4}DG [16]	–	$\underline{71.15}_{\pm 1.3}$	$\underline{70.34}_{\pm 2.2}$	$\mathbf{76.28}_{\pm \mathbf{1.8}}$	$58.46_{\pm 2.6}$	$\underline{69.06}_{\pm 1.9}$
Ours	–	$\mathbf{74.56}_{\pm \mathbf{1.0}}$	$\mathbf{73.01}_{\pm \mathbf{1.5}}$	$\underline{75.05}_{\pm 1.2}$	$\mathbf{62.58}_{\pm \mathbf{1.8}}$	$\mathbf{71.30}^{\dagger}_{\pm \mathbf{1.3}}$
Source Domain: C_1						
ERM	$61.87_{\pm 2.1}$	–	$59.45_{\pm 2.5}$	$63.01_{\pm 1.9}$	$50.67_{\pm 2.7}$	$58.75_{\pm 2.3}$
AdverIN [22]	$65.42_{\pm 1.8}$	–	$61.72_{\pm 2.4}$	$65.94_{\pm 1.6}$	$53.85_{\pm 2.5}$	$61.73_{\pm 2.1}$
ConDiSR [14]	$64.54_{\pm 1.9}$	–	$60.70_{\pm 2.3}$	$64.83_{\pm 1.7}$	$61.00_{\pm 2.2}$	$62.77_{\pm 2.0}$
RAS^{4}DG [16]	$\underline{73.39}_{\pm 1.5}$	–	$\underline{73.05}_{\pm 1.8}$	$\underline{76.48}_{\pm 1.4}$	$\mathbf{65.49}_{\pm \mathbf{2.0}}$	$\underline{72.10}_{\pm 1.7}$
Ours	$\mathbf{78.26}_{\pm \mathbf{1.2}}$	–	$\mathbf{75.20}_{\pm \mathbf{1.4}}$	$\mathbf{79.70}_{\pm \mathbf{1.1}}$	$\underline{64.08}_{\pm 1.6}$	$\mathbf{74.31}^{\dagger}_{\pm \mathbf{1.3}}$
Source Domain: C_2						
ERM	$55.12_{\pm 2.6}$	$47.46_{\pm 2.9}$	–	$51.09_{\pm 2.4}$	$54.65_{\pm 2.2}$	$52.08_{\pm 2.5}$
AdverIN [22]	$57.76_{\pm 2.3}$	$51.36_{\pm 2.7}$	–	$53.20_{\pm 2.1}$	$58.02_{\pm 2.0}$	$55.09_{\pm 2.3}$
ConDiSR [14]	$56.10_{\pm 2.4}$	$51.91_{\pm 2.6}$	–	$52.94_{\pm 2.2}$	$57.34_{\pm 2.1}$	$54.57_{\pm 2.4}$
RAS^{4}DG [16]	$\underline{75.71}_{\pm 1.6}$	$\mathbf{69.40}_{\pm \mathbf{1.9}}$	–	$\underline{68.35}_{\pm 1.7}$	$\underline{74.86}_{\pm 1.5}$	$\underline{72.08}_{\pm 1.7}$
Ours	$\mathbf{78.65}_{\pm \mathbf{1.3}}$	$\underline{67.72}_{\pm 1.8}$	–	$\mathbf{72.90}_{\pm \mathbf{1.4}}$	$\mathbf{77.97}_{\pm \mathbf{1.2}}$	$\mathbf{74.31}^{\dagger}_{\pm \mathbf{1.4}}$
Source Domain: C_3						
ERM	$53.49_{\pm 2.5}$	$46.45_{\pm 2.8}$	$54.41_{\pm 2.3}$	–	$38.25_{\pm 2.9}$	$48.15_{\pm 2.6}$
AdverIN [22]	$55.54_{\pm 2.2}$	$50.08_{\pm 2.5}$	$57.19_{\pm 2.0}$	–	$41.47_{\pm 2.7}$	$51.07_{\pm 2.4}$
ConDiSR [14]	$57.11_{\pm 2.1}$	$49.09_{\pm 2.6}$	$56.44_{\pm 2.1}$	–	$\underline{45.63}_{\pm 2.5}$	$52.07_{\pm 2.3}$
RAS^{4}DG [16]	$\underline{64.40}_{\pm 1.8}$	$\underline{53.76}_{\pm 2.4}$	$\mathbf{68.73}_{\pm \mathbf{1.7}}$	–	$43.98_{\pm 2.6}$	$\underline{57.72}_{\pm 2.1}$
Ours	$\mathbf{66.63}_{\pm \mathbf{1.5}}$	$\mathbf{57.86}_{\pm \mathbf{1.9}}$	$\underline{67.78}_{\pm 1.4}$	–	$\mathbf{47.65}_{\pm \mathbf{2.2}}$	$\mathbf{59.98}_{\pm \mathbf{1.8}}$
Source Domain: C_4						
ERM	$56.65_{\pm 2.4}$	$53.12_{\pm 2.5}$	$55.26_{\pm 2.3}$	$57.53_{\pm 2.2}$	–	$55.64_{\pm 2.4}$
AdverIN [22]	$\underline{59.56}_{\pm 2.1}$	$56.18_{\pm 2.2}$	$\underline{57.73}_{\pm 2.0}$	$\mathbf{61.42}_{\pm \mathbf{1.9}}$	–	$\mathbf{58.72}_{\pm \mathbf{2.0}}$
ConDiSR [14]	$59.35_{\pm 2.2}$	$\underline{56.74}_{\pm 2.3}$	$56.76_{\pm 2.1}$	$\underline{60.98}_{\pm 1.9}$	–	$58.46_{\pm 2.1}$
RAS^{4}DG [16]	$56.18_{\pm 2.3}$	$\mathbf{57.10}_{\pm \mathbf{2.1}}$	$56.17_{\pm 2.2}$	$56.28_{\pm 2.3}$	–	$56.43_{\pm 2.2}$
Ours	$\mathbf{59.70}_{\pm \mathbf{1.7}}$	$55.98_{\pm 1.9}$	$\mathbf{58.24}_{\pm \mathbf{1.8}}$	$60.64_{\pm 1.6}$	–	$\underline{58.64}_{\pm 1.8}$

(b) Diabetic Retinopathy

Method	Aptos	EyePP	Mess	Mess2	AVG.
Source: Aptos					
ERM	–	$42.69_{\pm 2.5}$	$30.32_{\pm 2.9}$	$38.11_{\pm 2.6}$	$37.04_{\pm 2.7}$
AdverIN [22]	–	$45.03_{\pm 2.3}$	$31.98_{\pm 2.8}$	$40.20_{\pm 2.4}$	$39.07_{\pm 2.5}$
ConDiSR [14]	–	$45.26_{\pm 2.2}$	$32.14_{\pm 2.7}$	$40.40_{\pm 2.3}$	$39.27_{\pm 2.4}$
RAS^{4}DG [16]	–	$\underline{48.14}_{\pm 2.0}$	$\underline{33.60}_{\pm 2.6}$	$\underline{42.75}_{\pm 2.2}$	$\underline{41.50}_{\pm 2.3}$
Ours	–	$\mathbf{51.85}_{\pm \mathbf{1.8}}$	$\mathbf{36.82}_{\pm \mathbf{2.1}}$	$\mathbf{46.28}_{\pm \mathbf{1.9}}$	$\mathbf{44.99}^{\dagger}_{\pm \mathbf{1.9}}$
Source: EyePACS					
ERM	$39.42_{\pm 2.7}$	–	$32.48_{\pm 2.8}$	$39.48_{\pm 2.5}$	$37.13_{\pm 2.7}$
AdverIN [22]	$41.58_{\pm 2.5}$	–	$34.26_{\pm 2.6}$	$41.65_{\pm 2.3}$	$39.16_{\pm 2.5}$
ConDiSR [14]	$41.79_{\pm 2.4}$	–	$34.43_{\pm 2.7}$	$41.86_{\pm 2.4}$	$39.36_{\pm 2.5}$
RAS^{4}DG [16]	$\underline{44.29}_{\pm 2.2}$	–	$\underline{36.14}_{\pm 2.5}$	$\underline{44.37}_{\pm 2.1}$	$\underline{41.60}_{\pm 2.3}$
Ours	$\mathbf{47.87}_{\pm \mathbf{1.9}}$	–	$\mathbf{39.44}_{\pm \mathbf{2.2}}$	$\mathbf{47.95}_{\pm \mathbf{1.8}}$	$\mathbf{45.09}^{\dagger}_{\pm \mathbf{2.0}}$
Source: Messidor					
ERM	$41.91_{\pm 2.4}$	$27.50_{\pm 3.0}$	–	$39.03_{\pm 2.6}$	$36.15_{\pm 2.7}$
AdverIN [22]	$44.20_{\pm 2.3}$	$29.01_{\pm 2.9}$	–	$41.17_{\pm 2.4}$	$38.13_{\pm 2.5}$
ConDiSR [14]	$44.43_{\pm 2.2}$	$29.16_{\pm 2.8}$	–	$41.37_{\pm 2.3}$	$38.32_{\pm 2.4}$
RAS^{4}DG [16]	$\underline{47.21}_{\pm 2.0}$	$\underline{30.30}_{\pm 2.7}$	–	$\underline{43.83}_{\pm 2.2}$	$\underline{40.45}_{\pm 2.3}$
Ours	$\mathbf{50.90}_{\pm \mathbf{1.7}}$	$\mathbf{33.40}_{\pm \mathbf{2.4}}$	–	$\mathbf{47.40}_{\pm \mathbf{1.9}}$	$\mathbf{43.90}^{\dagger}_{\pm \mathbf{2.0}}$
Source: Messidor-2					
ERM	$40.73_{\pm 2.5}$	$25.47_{\pm 3.0}$	$40.66_{\pm 2.5}$	–	$35.62_{\pm 2.7}$
AdverIN [22]	$42.96_{\pm 2.3}$	$26.87_{\pm 2.9}$	$42.89_{\pm 2.4}$	–	$37.57_{\pm 2.5}$
ConDiSR [14]	$43.18_{\pm 2.2}$	$27.00_{\pm 2.9}$	$43.11_{\pm 2.3}$	–	$37.76_{\pm 2.5}$
RAS^{4}DG [16]	$\underline{45.83}_{\pm 2.1}$	$\underline{27.91}_{\pm 2.8}$	$\underline{45.75}_{\pm 2.2}$	–	$\underline{39.83}_{\pm 2.4}$
Ours	$\mathbf{49.47}_{\pm \mathbf{1.8}}$	$\mathbf{30.94}_{\pm \mathbf{2.5}}$	$\mathbf{49.39}_{\pm \mathbf{1.9}}$	–	$\mathbf{43.26}^{\dagger}_{\pm \mathbf{2.1}}$

Implementation Details. *Backbones:* We primarily use DenseNet-121 initialized from scratch. For architectural ablation and stability analysis, we also report results using ResNet-18, ViT-Base, and ImageNet-pretrained DenseNet-121, where specified. *Training:* Models are trained for 30 epochs on a single NVIDIA A6000 GPU, which is sufficient for convergence under our schedule (validation accuracy typically plateaus by the final full-attack phase). We use a 5-epoch clean warm-up, followed by a linear ramp-up of the adversarial strength/budget until epoch 25, then 5 epochs of full attack. We use K=5 inner maximization steps (Algorithm 1), Weighting (λ) = 0.5, and a step size of 0.2, and study their sensitivity. Results are averaged over 5 fixed random seeds.

4 Results

Table 1 reports the most stringent setting, where training is restricted to a *single* source domain and evaluation spans all unseen domains. On Camelyon17-WILDS, PhaseAT delivers large, consistent gains over ERM and recent DG baselines, with the strongest improvements arising when the source center encourages the most pronounced center-specific shortcuts, for example, when trained on C_2, the average accuracy increases from 52.08 (ERM) to 74.31. Crucially, the

Table 2: Performance comparison in the MSDG setting. Reported values are the accuracy(%) mean ± standard deviation. Best and second-best per column are in **bold** and underline; † marks a significant gain over the strongest baseline (paired t-test, $p<0.05$).

(a) Camelyon17-WILDS

Method	C_0	C_1	C_2	C_3	C_4	AVG.
ERM	$78.89_{\pm 0.5}$	$84.57_{\pm 1.1}$	$67.24_{\pm 0.5}$	$84.47_{\pm 0.2}$	$74.08_{\pm 1.7}$	77.9
AdverIN [22]	$81.52_{\pm 0.6}$	$87.78_{\pm 1.0}$	$68.97_{\pm 0.7}$	$88.17_{\pm 0.4}$	$76.19_{\pm 1.6}$	80.5
ConDiSR [14]	$81.67_{\pm 0.4}$	$87.90_{\pm 0.7}$	$\underline{72.12}_{\pm 0.2}$	$86.79_{\pm 0.3}$	$\mathbf{80.06}_{\pm \mathbf{0.3}}$	$\underline{81.7}$
RAS^{4}DG [16]	$\underline{83.62}_{\pm 0.5}$	$\mathbf{89.10}_{\pm \mathbf{0.8}}$	$69.31_{\pm 0.4}$	$\underline{88.24}_{\pm 0.6}$	$77.60_{\pm 0.5}$	81.6
Ours	$\mathbf{90.49}_{\pm \mathbf{0.5}}$	$\underline{87.97}_{\pm 0.7}$	$\mathbf{76.86}_{\pm \mathbf{0.3}}$	$\mathbf{94.61}_{\pm \mathbf{0.5}}$	$\underline{78.58}_{\pm 0.4}$	**85.7**†

(b) Diabetic Retinopathy

Method	Aptos	EyePACS	Messidor	Messidor2	AVG.
ERM	$55.7_{\pm 1.2}$	$43.4_{\pm 0.8}$	$54.3_{\pm 1.6}$	$53.8_{\pm 0.4}$	51.8
AdverIN [22]	$58.6_{\pm 1.2}$	$48.0_{\pm 0.9}$	$58.3_{\pm 1.5}$	$53.4_{\pm 0.5}$	54.6
ConDiSR [14]	$59.4_{\pm 0.5}$	$48.8_{\pm 0.4}$	$59.2_{\pm 0.8}$	$54.2_{\pm 0.2}$	55.4
RAS^{4}DG [16]	$\underline{60.9}_{\pm 0.9}$	$\underline{50.2}_{\pm 0.6}$	$\underline{60.6}_{\pm 1.1}$	$\underline{55.6}_{\pm 0.3}$	$\underline{56.8}$
Ours	$\mathbf{62.4}_{\pm \mathbf{0.8}}$	$\mathbf{51.1}_{\pm \mathbf{0.5}}$	$\mathbf{62.1}_{\pm \mathbf{1.0}}$	$\mathbf{56.8}_{\pm \mathbf{0.3}}$	**58.1**†

gains are distributed across multiple target hospitals rather than concentrated on one, indicating reduced reliance on stain/texture-specific correlations and stronger sensitivity to morphology. On the Diabetic Retinopathy benchmark, PhaseAT similarly improves transfer across all source datasets (e.g., Aptos as source: 37.04 → 44.99 avg.), suggesting that enforcing stability to structured phase variations provides a robust inductive bias across heterogeneous acquisition conditions (camera/sensor, illumination, and preprocessing). Over the five shared seeds, gains marked † are significant against the strongest baseline (paired t-test, $p < 0.05$; e.g., SDG C_0: Δ=+2.24, 95% CI $[0.3, 4.1]$).

Following the SDG results in Table 1, Table 2 shows that the same phase-aware robustness mechanism remains effective when multiple source domains are available, albeit with a different gain regime. In SDG, improvements are often larger because single-source training amplifies center/dataset-specific shortcuts. In MSDG, pooling multiple sources already exposes the model to broader appearance variability, so the incremental gains are smaller but more stable. On Camelyon17-WILDS, PhaseAT achieves the best overall average (85.7%), substantially improving over ERM (77.9%) and surpassing recent DG methods clustered around $\sim$ 81–82%, with powerful benefits on shift-sensitive held-out hospitals (notably C_2), thereby improving worst-case behavior across centers. On Diabetic Retinopathy, PhaseAT likewise provides consistent improvements across all held-out datasets, increasing the average from 51.8% to 58.1% and outperforming the strongest baseline (RAS^{4}DG at 56.8%), indicating that enforcing stability to phase-driven structural perturbations complements multi-source supervision by discouraging reliance on dataset-specific color/contrast statistics and promoting geometry-aligned cues that transfer across acquisition pipelines.

Ablation Study. Figure 2 highlights both robustness and practicality. PhaseAT is comparatively stable as the batch size increases, consistent with the phase-perturbed supervision acting as an additional regularizer rather than relying on fragile optimization conditions. Performance is also insensitive to small variations in the loss weighting, with a broad optimum around $\lambda \approx 0.5$, indicating that the clean/phase-perturbed trade-off is not finely tuned. Moderate phase step sizes are beneficial, while overly many inner steps eventually degrade performance,

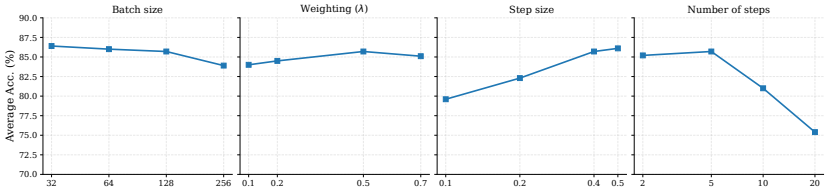

Fig. 2: Ablation studies for the main hyperparameters on the average accuracy over the 5 centers in MSDG setting on Camelyon17-WILDS.

Table 3: Ablation Study on Camelyon17-WILDS in MSDG setting.

(a) Backbone Performance

	C_0	C_1	C_2	C_3	C_4	Avg.
ERM_ResNet	78.83	83.56	79.00	90.66	60.39	78.49
Ours_ResNet	91.20	86.16	81.53	91.66	62.77	82.66
ERM_ViT	74.20	78.40	62.10	86.50	45.30	69.30
Ours_ViT	84.10	81.20	66.50	89.10	50.40	74.26
Densenet Imagenet Pretrained						
ERM	94.48	90.45	82.56	93.43	92.55	90.69
Ours	95.39	91.52	88.80	95.68	86.90	91.66

(b) Component Ablation

	C_0	C_1	C_2	C_3	C_4	Avg.
Amplitude-only	79.13	84.73	65.95	84.40	75.23	77.89
Phase+Amplitude	80.30	85.46	66.12	84.52	76.12	78.50
No Mask	60.01	81.68	55.82	49.93	34.92	56.47
RGB	85.61	85.65	61.21	86.01	74.62	78.62
YCbCr-all	79.86	86.14	73.18	90.54	77.64	81.47
Random Phase	89.57	87.87	64.01	91.82	76.22	81.90

consistent with perturbations becoming too strong and drifting away from the desired "style-controlled" stress test. Overall, the method exhibits a wide stable operating region and achieves its best results with moderate perturbation strength and a small number of steps.

Table 3(a) demonstrates architectural generality: PhaseAT improves ResNet-18 (78.49 → 82.66 avg.) and ViT-Base (69.30 → 74.26 avg.), indicating that the benefit is not tied to convolutional texture priors and persists for transformer backbones. With ImageNet-pretrained DenseNet, gains are smaller (90.69 → 91.66), suggesting that generic pretraining already contributes partial robustness, while PhaseAT still adds complementary invariances.

Table 3 (b) clarifies which choices matter. PhaseAT perturbs *phase* while *fixing* amplitude; when we instead (i) fix phase and update amplitude only or (ii) update both phase and amplitude (no amplitude fixing), performance drops, consistent with the intended "structure under matched style" shift. The saliency mask is particularly important: without it, performance degrades sharply, indicating that concentrating updates on phase-sensitive frequencies stabilizes training. Operating on luminance is consistently advantageous: RGB-space perturbations degrade performance, and perturbing all YCbCr channels recovers only partially, reinforcing the choice to emphasize geometry and avoid unnecessary chromatic distortions. The "Random Phase" variant (sampling a bounded random phase offset within the same ℓ_∞ budget, instead of gradient-based inner

maximization) remains strong, showing that phase jitter is a useful regularizer, but the gradient-guided procedure provides the best overall generalization.

5 Conclusion & Future Work

We introduce PhaseAT, a phase-aware adversarial training framework that perturbs the Fourier phase of the luminance channel while fixing amplitude, generating structure-stressing views with matched appearance and reducing reliance on domain-specific texture shortcuts. PhaseAT improves transfer on Camelyon17-WILDS and diabetic retinopathy, achieving up to +22% accuracy gains over ERM in single-source settings and +9% multi-source. Our evaluation is currently limited to 2D classification and color medical images (YCbCr); investigating performance on grayscale modalities remains an important limitation. Future work will extend phase-constrained adversarial views to dense prediction, 3D modalities, and broader acquisition shifts, and study more adaptive frequency masking/selection, combinations with complementary amplitude-/style-based DG objectives, and efficiency improvements that reduce training overhead while preserving robustness.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. MESSIDOR-2 dataset. ADCIS (dataset access page), https://www.adcis.net/en/third-party/messidor2/, accessed: 2026-01-27
2. Diabetic retinopathy detection. Kaggle Competition (2015), https://www.kaggle.com/c/diabetic-retinopathy-detection, accessed: 2026-01-27
3. APTOS 2019 blindness detection. Kaggle Competition (2019), https://www.kaggle.com/c/aptos2019-blindness-detection, accessed: 2026-01-27
4. Bándi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., Li, Q., Zanjani, F.G., Zinger, S., 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 (2019). https://doi.org/10.1109/TMI.2018.2867350
5. Chen, C., Li, Z., Ouyang, C., Sinclair, M., Bai, W., Rueckert, D.: Maxstyle: Adversarial style composition for robust medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Lecture Notes in Computer Science, vol. 13435, pp. 151–161 (2022). https://doi.org/10.1007/978-3-031-16443-9_15
6. Chen, X., Shen, Z., Cao, P., Yang, J., Zaiane, O.R.: Constyx: Content style augmentation for generalizable medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. Lecture Notes in Computer Science, vol. 15964, pp. 100–110 (2025). https://doi.org/10.1007/978-3-032-04971-1_10

7. Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., Charton, B., Klein, J.C.: Feedback on a publicly distributed image database: The MESSIDOR database. Image Analysis & Stereology **33**(3), 231–234 (2014). https://doi.org/10.5566/ias.1155
8. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations (ICLR) (2019), https://arxiv.org/abs/1811.12231, iCLR 2019 (oral)
9. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization (2020), https://arxiv.org/abs/2007.01434
10. Hu, D., Li, H., Liu, H., Yao, X., Wang, J., Oguz, I.: Map: Domain generalization via meta-learning on anatomy-consistent pseudo-modalities. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops (MedAGI). Lecture Notes in Computer Science, vol. 14393. Springer (2023). https://doi.org/10.1007/978-3-031-47401-9_18
11. Jiang, J., Gu, S.: Train once, deploy anywhere: Edge-guided single-source domain generalization for medical image segmentation. In: Burgos, N., Petitjean, C., Vakalopoulou, M., Christodoulidis, S., Coupe, P., Delingette, H., Lartizien, C., Mateus, D. (eds.) Proceedings of The 7nd International Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 250, pp. 722–741. PMLR (03–05 Jul 2024), https://proceedings.mlr.press/v250/jiang24a.html
12. Koor, K., Qiu, Y., Kwek, L.C., Rebentrost, P.: A short tutorial on wirtinger calculus with applications in quantum information (2023), https://arxiv.org/abs/2312.04858
13. Lee, S., Bae, J., Kim, H.Y.: Decompose, adjust, compose: Effective normalization by playing with frequency for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11776–11785 (2023). https://doi.org/10.1109/CVPR52729.2023.01134
14. Matsun, A., Saeed, N., Maani, F.A., Yaqub, M.: Condisr: Contrastive disentanglement and style regularization for single domain generalization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2881–2889. IEEE (2025)
15. Pestana, C., Akhtar, N., Liu, W., Glance, D., Mian, A.: Adversarial attacks and defense on deep learning classification models using ycbcr color images. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–9 (2021). https://doi.org/10.1109/IJCNN52387.2021.9533495
16. Qiao, Q., Wang, W., Qu, M., Su, K., Jiang, B., Guo, Q.: Medical image segmentation via single-source domain generalization with random amplitude spectrum synthesis. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer Science, vol. 15009, pp. 435–445. Springer (2024). https://doi.org/10.1007/978-3-031-72114-4_42
17. Safdari, R., Nikouei Mahani, M.A., Koohi-Moghadam, M., Bae, K.T.: MixStyle-Flow: Domain Generalization in Medical Image Segmentation using Normalizing Flows. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. Lecture Notes in Computer Science, vol. 15962, pp. 376–385 (2025). https://doi.org/10.1007/978-3-032-04947-6_36
18. Su, Z., Yao, K., Yang, X., Wang, Q., Sun, J., Huang, K.: Rethinking data augmentation for single-source domain generalization in medical image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2366–2374 (2023). https://doi.org/10.1609/aaai.v37i2.25332

19. Teterwak, P., Saito, K., Tsiligkaridis, T., Plummer, B.A., Saenko, K.: Is large-scale pretraining the secret to good domain generalization? In: Proceedings of the International Conference on Learning Representations (ICLR) (2025), https://openreview.net/forum?id=wCOJpXm0Me
20. Yi, J., Bi, Q., Zheng, H., Zhan, H., Ji, W., Huang, Y., Li, S., Li, Y., Zheng, Y., Huang, F.: Hallucinated style distillation for single domain generalization in medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer Science, vol. 15010, pp. 438–448. Springer (2024). https://doi.org/10.1007/978-3-031-72117-5_41
21. Zhang, L., Luo, Y., Shen, H., Wang, T.: A fourier perspective of feature extraction and adversarial robustness. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI). pp. 1715–1723 (2024). https://doi.org/10.24963/ijcai.2024/190
22. Zhang, Z., Wang, B., Yao, L., Keles, E., Jha, D., Antalek, M., Durak, G., Medetalibeyoglu, A., Spampinato, C., Turkbey, B., Gong, B., Bagci, U.: Adverin: Monotonic adversarial intensity attack for domain generalization in medical image segmentation. Medical Image Analysis **107**, 103848 (2026). https://doi.org/10.1016/j.media.2025.103848
23. Zhang, Z., Chen, N., Liu, Y., Zhang, Q., Zhang, X.: Adversarial data augmentation for single domain generalization via lyapunov exponent-guided optimization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025), https://arxiv.org/abs/2507.04302
24. Zhou, D., Wang, N., Yang, H., Gao, X., Liu, T.: Phase-aware adversarial defense for improving adversarial robustness. In: Proceedings of the 40th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 202, pp. 42724–42741 (2023)