GS-MIX: SIMPLE PIXEL-LEVEL MIXING FOR SINGLE-DOMAIN GENERALIZED OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Pixel-level data augmentation has proven to be an effective strategy for single-domain generalized object detection (S-DGOD). However, existing methods struggle to balance augmentation intensity with the preservation of object-relevant structures, thereby restricting strong perturbations for diverse coverage. To address this, we analyze the key factors and influence mechanism in S-DGOD domain shift, and accordingly propose a vicinal distribution that synthesizes samples with domain shift perturbations (noise and texture) while robustly preserving structure. To instantiate this distribution, we construct a simple yet effective pixel-level intra-domain mixing framework, termed Global-context Structure-texture Mix (GS-Mix). Specifically, global-context mixing (GCM) yields diverse noise perturbations through linear interpolation between source images, and structure-texture mixing (STM) uses structure masks to apply strong texture transformations. These modules cover global-to-local strong augmentations and avoid destroying structural information. GS-Mix can be easily taken as a plug-and-play data augmentation module and introduces almost no extra costs in the training phase. Experimental results on real-world domain generalization benchmarks, including adverse weather, underwater, and medical imaging, show the superior performance and easy scalability of our GS-Mix compared to existing methods.

1 INTRODUCTION

Object detection aims to precisely locate and recognize specific objects within images. It has been widely applied in various systems equipped with visual perception modules, such as autonomous driving, medical diagnosis (Sultani et al., 2022), and underwater operations (Luo et al., 2025). However, in real-world scenarios, the deployment environments for detectors often exhibit data distributions that are inconsistent with the training domain, such as the transition from normal to adverse weather conditions (Michaelis et al., 2019). The resulting domain shift problem inevitably leads to severe performance degradation of the detectors.

Collecting and annotating images from multiple domains is time-consuming and labor-intensive, which hinders the flexible deployment of detectors. Therefore, domain adaptation (DA) and single domain generalization (SDG) have been proposed to extend detectors to target domains with different distributions without relying on full target data. DA methods utilize unlabeled target images to mitigate domain shift and can be categorized into domain alignment (Chen et al., 2018; 2020; Wen et al., 2025), domain translation (Kim et al., 2019; Hsu et al., 2020), and self-training methods (Chen et al., 2022; Pu et al., 2024). Although these methods can significantly improve detector performance in the target domain, they are limited by their reliance on target domain images; pre-collecting such images is difficult for certain scenarios, such as rare weather conditions.

Consequently, single-domain generalized object detection (S-DGOD) has recently gained significant attention (Wang et al., 2023; Li et al., 2026). Its goal is to generalize a detector trained on a single labeled source domain to unseen target domains, without access to target data during training. S-DGOD methods can be categorized by paradigm into domain-invariant strategies, pretrained foundation priors, and data augmentation techniques. Domain-invariant strategies (Wu & Deng, 2022; Wu et al., 2024) aim to learn invariant features from source data via consistency learning or domain disentanglement. However, single invariance mechanism often struggles to bridge the multi-domain gap, resulting in limited gains. Pre-trained foundation prior methods (Vidit et al., 2023; Jiang et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2025; He et al., 2025a) exploit knowledge from pre-trained foundation models on large-scale general data. While this approach brings effective and diversified information, it often introduces significant additional training overhead.

Data augmentation-based S-DGOD approaches expand the training distribution by generating vicinal samples via perturbations on source data. These methods can be broadly categorized into feature-level and input-level augmentation. Feature-level augmentation (Jiang et al., 2025; Wu & Deng, 2025) perturb intermediate network representations, which may inadvertently alter object-relevant semantic information. Input-level augmentation operates on the input images and can be further divided into frequency-domain and pixel-level methods. While frequency-domain methods (Xu et al., 2025) operate in the spectral space, they often suffer from restricted augmentation directions and the complexity of domain transformations. In contrast, pixel-level augmentation (Danish et al., 2024; Lee et al., 2024; Liu et al., 2024) directly manipulates the image space, which is easier to implement and naturally provides a wide range of controllable perturbation directions, making it an effective and lower-cost approach.

However, existing pixel-level augmentation methods still exhibit notable limitations: 1) They typically rely on auxiliary domain-invariant training strategies (e.g., consistency learning) to ensure performance. This dependency increases framework complexity and hinders their portability across diverse model architectures; 2) They fail to effectively balance augmentation intensity with the preservation of object-relevant structure. In practice, they often reduce perturbation intensity in object regions to avoid damaging object structure, thereby limiting the coverage of strong perturbations that mimic domain shifts.

To address these limitations, grounded in the vicinal risk minimization (VRM) paradigm (Chapelle et al., 2000), we analyze the domain shift problem in real-world S-DGOD and propose an object-centric VRM distribution. This distribution focuses on synthesizing vicinal samples that cover critical domain shift factors (i.e., noise perturbation and texture transformation) while preserving domain-invariant elements (i.e., object-relevant structures). Accordingly, we propose a pure pixel-level data augmentation framework to instantiate this distribution, called **G**lobal-context **S**tructure-texture **M**ix (GS-Mix). Specifically, we introduce global-context mixing (GCM) to synthesize diverse noise-perturbed samples via direct intra-domain pixel-wise mixing. Additionally, we propose structure-texture mixing (STM), a module that mixes structure maps with texture-augmented images via structure masks, thereby yielding structure-preserving texture perturbations. Experimental results on S-DGOD tasks demonstrate the effectiveness of our proposed method.

Compared to existing pixel-level augmentation methods, our GS-Mix does not rely on carefully searching for specific corruption combinations to extend data distribution, and it maintains the detector's sensitivity to structure while augmenting source image. This design can naturally address the key domain shift challenges of noise perturbations and texture transformations in S-DGOD, driving the detector toward reliable VRM learning.

Notably, GS-Mix operates entirely in the pixel-level, without requiring transformation to the frequency domain or operations at the feature-level. This means it can be simply plugged into the data loading stage as an augmentation component. Compared to other domain-invariant strategies and pre-trained foundation prior methods, our GS-Mix offers great convenience with an almost negligible additional training cost. Moreover, it delivers competitive performance on popular lightweight one-stage detectors (e.g., the YOLO series), demonstrating exceptional practical value.

Our contributions are summarized as follows:

1. We analyze S-DGOD domain shift factors and mechanisms and propose an object-centric VRM distribution. This distribution is designed to synthesize samples that cover critical domain shift factors while preserving domain-invariant elements.
2. To instantiate this distribution, we construct two streamlined pixel-level operation components: global-context mixing to synthesize noise-perturbed samples, and structure-texture mixing to generate structure-augmented samples. This approach covers vicinal samples from global-to-local levels through simple mixing operations.
3. Extensive experiments on S-DGOD benchmarks demonstrate that our method substantially outperforms previous methods. Furthermore, we establish a benchmark choice for lightweight detectors, achieving an optimal trade-off between accuracy and inference speed.

2 RELATED WORKS

Data Augmentation for Generalization. Data augmentation expands training samples through predefined transformation functions, representing one of the cheapest and simplest methods to improve model domain generalization (Wang et al., 2023). Conventional approaches (Honarvar Nazari & Kovashka, 2020) typically apply linear transformations on original images, such as rotation and translation. Established works like AutoAugment (Cubuk et al., 2018) employ reinforcement learning to search the transformation space for optimal deployment; these methods are applicable to domain generalization tasks. Beyond geometric shifts, domain randomization techniques (Tobin et al., 2017; Zhao et al., 2023) simulate diverse distributions by manipulating lighting, injecting noise, or composing virtual objects.

Recently, data generation based on the linear combination of samples has gained attention due to its high efficiency. As an early idea, Mixup (Zhang et al., 2018) shows that linearly interpolating any two source domain images to create new samples can promote model generalization. Subsequently, numerous improved variants have been proposed and validated in domain generalization tasks. For example, AugMix (Hendrycks et al., 2019) mixes multiple augmented views of the same image, and C-Mixup (Yao et al., 2022) adjusts the sampling probability for mixing according to label similarity. EDM (Cao & Chen, 2024) extends Mixup to an extrapolation domain strategy. DomCLP (Lee et al., 2025) performs mixup on samples from different domains using cluster centroids as supervision. Some methods further extend mixing to the feature-level, e.g., CuMix (Mancini et al., 2020) that mixes color patches with feature information, and MixStyle (Zhou et al., 2024), which mixes feature statistics across different images. Despite the widespread adoption of intra-domain mixing, its efficacy remains constrained for S-DGOD, which requires balancing domain shift mitigation with precise multi-object recognition. Therefore, leveraging the inherent efficiency of pixel-level mixing, we further investigate its potential for S-DGOD.

Domain Generalization Object Detection. S-DGOD aims to generalize a detector trained on a single labeled source domain to unseen domains. CDSD (Wu & Deng, 2022) first proposed the S-DGOD task, employing domain disentanglement to learn representations of domain-invariant features. Following a similar objective, G-NAS (Wu et al., 2024) utilizes neural architecture search and designs a generalization loss to capture object-relevant invariants, while SoMA (Yun et al., 2025) introduces singular value decomposition to preserve domain-invariant generalization capabilities via fine-tuning. Another paradigm leverages large-scale pre-trained foundation models. ClipGap (Vidit et al., 2023) utilizes pre-trained vision-language models with text prompts to inject diverse domain knowledge. SE-COT (Zhang et al., 2025) proposes a chain-of-thought method, which utilizes CLIP to progressively integrate and expand estimated cross-domain style information. GDD (He et al., 2025b) and FGT (He et al., 2025a) propose using pretrained diffusion models as strong foundation priors for knowledge transfer.

Some methods adopt a combination of data augmentation and additional training strategies. DivAlign (Danish et al., 2024) optimizes generalization via selected data augmentations and domain alignment. OA-DG (Lee et al., 2024) proposes combining object-aware mixing with a mixing loss, avoiding strong perturbations in object regions. UFR (Liu et al., 2024) leverages SAM to extract object regions for gentle local augmentation while applying frequency-domain augmentation to backgrounds, integrated with causal prototype learning. DiffDet (Jiang et al., 2025) performs feature augmentation by adding noise via a diffusion process and utilizing a pre-trained text encoder for denoising. Conversely, some methods focus exclusively on data augmentation exploration. PhysAug (Xu et al., 2025) introduces a physical model based augmentation that perturbs the frequency spectrum to simulate non-ideal imaging conditions (e.g., weather and light). Additionally, LossMix (Vu et al., 2024) investigates pixel-level augmentation for domain adaptation, though it necessitates access to target data. In contrast to these approaches, we propose a purely pixel-level augmentation framework for S-DGOD. Our method does not rely on complex frequency domain transformations or feature-level manipulations, offering a convenient, plug-and-play solution to enhance generalization performance.

3 METHOD

For the problem definition of S-DGOD, consider a source domain dataset $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{N}$, which consists of N images x_i and corresponding detection annotations y_i (comprising bounding boxes B

and categories C). Let $P_{\mathcal{S}}$ denote the data distribution of the source domain. The goal of S-DGOD is to train a model solely on a single source domain that can generalize to unseen target domains $\mathcal{T}$, where the target data distribution $P_{\mathcal{T}}$ differs from the source distribution $P_{\mathcal{S}}$. Addressing this problem is highly practical in real-world applications. For instance, in traffic scenarios, a detector trained on normal weather data often faces challenges from various unseen weather domains.

In standard supervised training on $\mathcal{S}$, the empirical risk minimization (ERM) objective, $\min_\theta \frac{1}{N}\sum \mathcal{L}(f(x_i), y_i)$, tends to approximate the distribution $P'_{\mathcal{S}}$ as a sum of Dirac δ functions over the training samples. This tendency exacerbates object-irrelevant biases inherent in the source domain, rendering the model difficult to circumvent domain shift problems. Conversely, data augmentation, formalized by the principle of vicinal risk minimization (VRM), leverages human-defined heuristics to synthesize vicinal samples. This expands the support of the training set and can improve the model's domain generalization capability. However, it has limitations in S-DGOD:

1) Classical VRM typically constructs an isotropic vicinal distribution, assuming that small perturbations in all directions are safe. This does not align with reality, where real domain shifts exist only in a few directions with physical or semantic significance.
2) Unlike image classification, object detection is a multi-task learning problem. It requires the model not only to be robust against domain shift elements but also to stably recognize edges and shapes for category classification and bounding box regression. Blind data augmentation that destroys object-relevant elements, such as strong masking and shape destruction, directly leads to negative performance impacts.

Therefore, we analyze the domain shift in real-world scenarios. Based on established priors and existing research (Wang et al., 2023; Oza et al., 2024), we focus on representative and dominant domain shift factors:

1) **Noise perturbation shift** η, such as variations in weather and illumination, which introduces external noise interference relative to the source.
2) **Texture transformation shift** t, including alterations in object material and artistic style, which essentially amplifies object-irrelevant features.

In contrast to these appearance shifts, the **object-relevant structural elements** u (e.g., shape or geometry) are expected to remain largely consistent across domains to support reliable recognition. For an image x from a potentially unseen domain, we model its formation as a composition of these underlying factors:

$$x = g(\eta, t, u), \tag{1}$$

where g denotes the composition function that interacts these factors to form the visual observation.

Building on this, we conduct object-centric VRM (OC-VRM) as the principle for S-DGOD. The core idea is to establish an anisotropic vicinity within the input space that spans along the directions of texture t and noise η variations, while maintaining invariance in the structural dimension u. We formulate the vicinal distribution P_v as a mixture of complementary directions:

$$P_v(\tilde{x}, \tilde{y} \mid x_i, y_i, \mathbf{c}) = \pi \cdot P_\eta(\tilde{x}, \tilde{y} \mid x_i, y_i, \mathbf{c}) + (1-\pi) \cdot P_t(\tilde{x}, \tilde{y} \mid x_i, y_i, \mathbf{c}, \mathbf{m}), \tag{2}$$

where $\pi \in [0, 1]$ represents the mixing coefficient. Here, $\mathbf{c}$ denotes the global context derived from the source domain distribution $P_{\mathcal{S}}$. The term P_t synthesizes samples by perturbing the texture factor t using information from $\mathbf{c}$, while adhering to a structural constraint $\mathbf{m}$ intrinsic to x_i to preserve the u. Complementarily, P_η utilizes $\mathbf{c}$ to inject global noise perturbations η without compromising structural integrity. Crucially, these stochastic perturbations extend the vicinal support to cover potential variations of nuisance factors, retaining the object-relevant elements to mitigate overfitting to the domain shift factors.

The objective is to minimize the vicinal risk $\mathcal{R}_{\text{OC-VRM}}(\theta)$ over this expanded support:

$$\mathcal{R}_{\text{OC-VRM}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathbf{c}\sim P_{\mathcal{S}}}\Big[\mathbb{E}_{(\tilde{x},\tilde{y})\sim P_v(\cdot|x_i,y_i,\mathbf{c})}[\mathcal{L}(f_\theta(\tilde{x}), \tilde{y})]\Big]. \tag{3}$$

The key lies not in acquiring more data, but in designing a vicinal distribution P_v that aligns closer with potential domain shifts $P_{\mathcal{T}}$. This ensures the training process covers texture transformations and noise perturbations, thereby compelling the detector to learn more stable structural discriminative cues to mitigate domain shift effects.

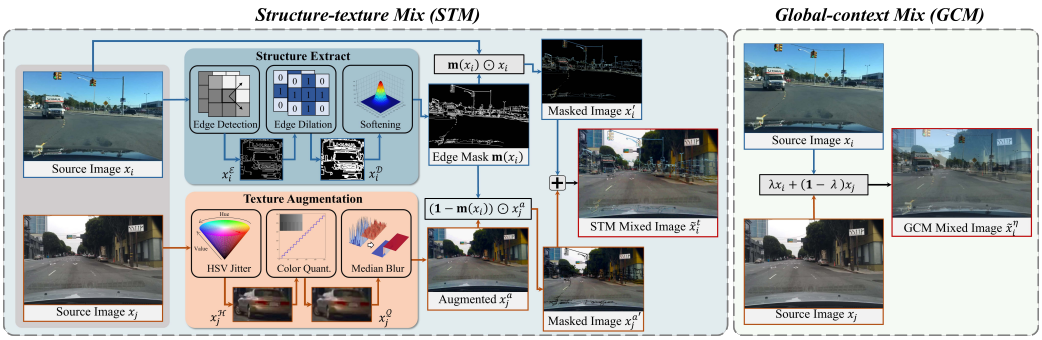

Figure 1: Overview of the GS-Mix pipeline, which is composed of STM and GCM. In STM, source images x_i and x_j undergo structure extraction (edge extraction, dilation, smoothing) to yield mask $\mathbf{m}(x_i)$, and texture augmentation (HSV jitter, quantization, median blurring) to yield x_j^a. Then, x_j^a and x_i are mixed using $\mathbf{m}(x_i)$ as the interpolation factor to generate the STM mixed image $\tilde{x}_i^t$. In GCM, x_i and x_j are mixed via linear interpolation to generate $\tilde{x}_i^\eta$. These modules are executed alternately during training.

This perspective affords OC-VRM significant extensibility. By flexibly tailoring components $\mathbf{c}$ and $\mathbf{m}$, OC-VRM can be instantiated in diverse forms. For instance, within conditional diffusion models (Yang et al., 2024; Zhang et al., 2023), where the condition serves as $\mathbf{m}$ and source domain knowledge acts as $\mathbf{c}$.

In this study, prioritizing methodological universality, we employ two simple pixel-level modules that serve as an efficient approximation of the OC-VRM: 1) We construct a GCM module, which samples different images from the source domain for linear group mixing to generate global and diverse noise perturbations. 2) We introduce a STM module designed to augment texture while preserving structure, allowing the detector to focus on structural information. The representation in the feature space is depicted in Fig. 2.

3.1 Realizing P_η: Global-context Mixing

To instantiate the noise perturbation vicinal distribution P_η within OC-VRM, we construct a direct and effective GCM module. GCM aims to synthesize global and structure-preserving noise vicinal samples. Specifically, we employ a Mixup (Zhang et al., 2018) operation between samples. Given two source domain samples x_i and x_j, where $j \in \{1, \ldots, N\} \setminus i$, we sample $\lambda \sim \text{Beta}(\alpha, \alpha)$ to construct the virtual vicinity:

$$\tilde{x}_i^\eta = \lambda x_i + (1-\lambda) x_j, \quad \tilde{y}_i = y_i \cup y_j. \tag{4}$$

By performing linear interpolation on images, it reduces the effective signal-to-noise ratio and smooths decision boundaries, approximating environmental degradation. This introduces diverse noise perturbations without causing geometric transformations or structural destruction.

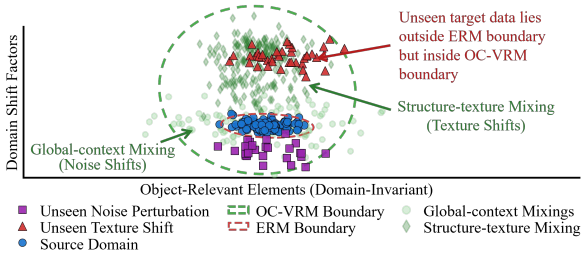

Figure 2: S-DGOD mechanisms in feature space. ERM is confined to the source domain and fails under domain shifts. Our OC-VRM extends the feature space via decoupled augmentations: 1) GCM constructs a vicinity covering unseen noise perturbations; 2) STM covers severe unseen texture shifts.

Within our OC-VRM, GCM serves as a baseline regularizer for noise perturbations. However, in S-DGOD tasks, besides resisting noise perturbations, the model must focus on extracting object-relevant elements (e.g., shape and edges) and avoid being affected by texture transformations. As evidenced by the comparative analysis in Fig. 3, while GCM provides substantial gains for scenarios dominated by noise perturbation (normal-to-adverse), it yields only marginal improvements for severe

texture shifts (real-to-cartoon). This is because, although linear mixing introduces diversity in global color and intensity statistics, it remains inherently strictly linear at the pixel-level. Consequently, it is insufficient to cover the complex variation directions characteristic of texture shifts.

3.2 Realizing P_t: Structure-texture Mixing

To address the limitations of GCM, we introduce STM as the instantiation of P_t. Using a similarly simple pixel-level intra-domain mixing approach, we construct vicinal samples that are structurally invariant yet texture-augmented. This aims to learn structure-prioritized discriminative cues and suppress texture bias.

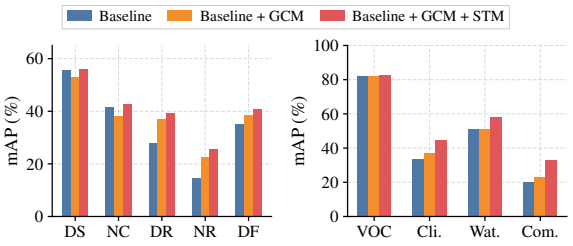

Figure 3: GCM effectively mitigates domain shift in domain generalization tasks dominated by noise perturbation (normal-to-adverse, *Left*). However, it provides limited gains for tasks dominated by texture transformation (real-to-cartoon, *Right*).

Formally, given source samples x_i and x_j, we first perform structure extraction on x_i. Edge data is obtained via Canny $\mathcal{E}$ and Dilation $\mathcal{D}$ operators. To prevent sharp edges in the subsequent mixing, we apply Gaussian blur as a smoothing function $\mathcal{G}$ to obtain the edge mask $\mathbf{m} \in [0,1]^{H\times W}$:

$$\mathbf{m}(x_i) = \mathcal{G}(\mathcal{D}(\mathcal{E}(x_i))). \tag{5}$$

Next, we perform texture augmentation on x_j. We explicitly restrict our design to an operation space that alters only texture and color statistics while maintaining geometric and semantic fidelity. This constraint is critical to prevent the induction of annotation noise arising from geometric distortion. Accordingly, we select three complementary axes: HSV jitter $\mathcal{H}$ (simulating color drift), color quantization $\mathcal{Q}$ (compressing color freedom and material details), and median blur $\mathcal{B}$ (suppressing high-frequency texture while preserving edges). This simple combination operation Ψ covers a broad range of appearance transformations:

$$x_j^a = \Psi(x_j) = \mathcal{B}(\mathcal{Q}(\mathcal{H}(x_j))). \tag{6}$$

Finally, using the softened edge mask $\mathbf{m}(x_i)$ (channel-expanded) as a smooth transition, we mix the x_i and the texture-augmented x_j^a to generate the vicinal sample $\tilde{x}_i^t$:

$$\tilde{x}_i^t = \mathbf{m}(x_i) \odot x_i + (1 - \mathbf{m}(x_i)) \odot x_j^a, \quad \tilde{y}_i = y_i \cup y_j. \tag{7}$$

Notably, in the generated image $\tilde{x}_i^t$, the augmented texture of x_j^a is mapped onto the object structure of x_i, constructing a texture-diversified distribution by constrained by original layouts. To minimize vicinal risk, the model need ignore texture t and focus on structure u as the discriminative cues.

GS-Mix Pipeline. The GS-Mix pipeline is shown in Fig. 1. From an implementation standpoint, we adopt a stochastic and mutually exclusive sampling strategy for vicinal sample generation during training. For each training sample x_i, we sample a selector variable $\gamma \sim \text{Bernoulli}(\pi)$. If $\gamma = 1$, we activate the GCM module to generate $\tilde{x}_i^\eta$ and skip STM. Otherwise, we only activate the STM module to generate $\tilde{x}_i^t$. The activation of each specific path is further governed by the probability of texture p_t and noise probability p_η. If no perturbation is triggered, an identity mapping is applied to ensure P_v remains normalized. By using this mutually exclusive mechanism, we avoid potential structural destruction caused by the superposition of strong augmentations, guiding the model to alternately and comprehensively cover the OC-VRM.

4 Experiments

4.1 Datasets

For the **Normal-to-adverse weather** task, by definition (Vidit et al., 2023), only labeled source domain data is used during training, which contains 19,395 images of normal weather and lighting

Table 1: Results of normal-to-adverse weather task. "$*$" indicates the backbone training parameters, and "†" represents inference latency, measured as the time to process a single 640×640 image on an identical single-GPU setup for both Faster R-CNN and YOLOv8s. GCM stands for our global context mixing module.

Methods	Detector	Paradigms	DS	NC	DR	NR	DF	mPC	Params (M)$*$	Speed (ms)†
FRCNN-R101	FRCNN	-	50.4	37.5	29.2	14.6	33.1	28.6		
CDSD (Wu & Deng, 2022)	FRCNN	Domain-Invariant Strategies	56.1	36.6	28.2	16.6	33.5	28.7		
G-NAS (Wu et al., 2024)	FRCNN		58.4	45.0	35.1	17.4	36.4	33.5		
SoMA (Yun et al., 2025)	FRCNN		49.3	41.9	37.9	24.5	38.2	35.6		
ClipGap (Vidit et al., 2023)	FRCNN	Pre-trained Foundation Priors	51.3	36.9	32.3	18.7	38.5	31.6		
SE-COT (Zhang et al., 2025)	FRCNN		55.4	42.0	39.2	24.5	40.6	36.6		
GDD (He et al., 2025b)	FRCNN		-	48.7	37.4	21.7	44.7	38.1		
FGT (He et al., 2025a)	FRCNN		-	**50.5**	39.1	22.4	**46.7**	39.7		
WFS (Wu & Deng, 2025)	FRCNN	Data Aug. (Feature)	-	38.5	34.1	20.2	40.1	33.2		
DiffDet (Jiang et al., 2025)	FRCNN		**61.8**	44.7	40.9	**26.6**	43.2	38.9	42.3	42.9
PhysAug (Xu et al., 2025)	FRCNN	Data Aug. (Fourier)	60.2	44.9	41.2	23.1	40.8	37.5		
OA-DG (Lee et al., 2024)	FRCNN	Data Aug. (Pixel) + Domain-Invariant Strategies	56.4	38.0	33.9	16.8	38.3	31.8		
UFR (Liu et al., 2024)	FRCNN		58.6	40.8	33.2	19.2	39.6	33.2		
DivAlign (Danish et al., 2024)	FRCNN		52.8	42.5	38.1	24.1	37.2	35.5		
GCM	FRCNN	Data Aug. (Pixel)	55.9	39.3	40.6	23.6	41.9	36.4		
GS-Mix	FRCNN	Data Aug. (Pixel)	59.8	46.9	**42.0**	26.4	45.9	**40.3**		
YOLOv8s	YOLOv8	-	55.8	41.7	27.9	14.6	35.3	29.9		
GCM	YOLOv8	Data Aug. (Pixel)	53.0	38.3	37.1	22.5	38.5	34.1	5.08	7.7
GS-Mix	YOLOv8	Data Aug. (Pixel)	**56.1**	**42.8**	**39.2**	**25.5**	**40.7**	**37.1**		

conditions. The source domain validation set Daytime Sunny (DS) includes 8,313 images, while testing is conducted on 4 target domain test sets: 2,494 images of Night Rainy (NR), 3,775 images of Daytime Foggy (DF), 3,501 images of Dusk Rainy (DR), and 26,158 images of Night Clear (NC). For the **Real-to-cartoon** task, consistent with the settings in previous work (Danish et al., 2024), we utilize 16,551 labeled real-world images from the train and val sets of VOC 2007 and VOC 2012 as the source training data. The unseen test sets comprise the test splits of Watercolor2k and Comic2k (1,000 images each) and all 1,000 images from Clipart1k. For the **Medical domain generalization** task, we adopt the M5 dataset for malaria detection across four developmental stages. Following previous work (Sultani et al., 2022), images from the high-cost microscope (HCM) serve as the source domain, while those from the low-cost microscope (LCM) serve as the target domain, with evaluation performed at matching magnifications ($1000\times$ and $400\times$). For the **Underwater domain generalization** task, we adopt S-UODAC2020, which covers seven domains and four object categories. Following previous work (Luo et al., 2025), the source training set contains 4,745 images from domains 1–6, while the target test set contains 797 images from domain 7.

4.2 Performance Analysis

In this section, we provide an experimental comparative analysis with current state-of-the-art S-DGOD methods. Additionally, we report the performance of GS-Mix on the lightweight YOLOv8s detector, highlighting its distinct advantages relative to the Faster R-CNN baseline. Notably, in the experiments employing YOLOv8s as the baseline for the underwater and medical generalization tasks, we directly apply the exact same parameters and configurations as those used in the normal-to-adverse weather task. This is to verify the ease of engineering reusability of our method.

Normal-to-adverse weather. In this comparison, we categorize existing S-DGOD methods based on their generalization paradigms, including domain-invariant strategies, pre-trained foundation priors, data augmentation, and hybrid methods. As presented in Tab. 1, methods utilizing pre-trained foundation priors (e.g., GDD, FGT) have achieved significant progress by introducing general knowledge. Notably, however, our GCM module using only simple pixel-level data augmentation achieved an mPC of 36.4%, approaching the performance of several recent methods. Our GS-Mix, which also relies solely on pixel-level intra-domain mixing, achieves the best mPC. GS-Mix outperforms DivAlign (an S-DGOD method that uses extensive data augmentation exploration and carefully chosen augmentation policies) by 4.8 mPC points. These results indicate that GS-Mix is a competitive alternative.

Table 2: Results of real-to-cartoon task.

Methods	Detector	VOC	Cli.	Wat.	Com.	mPC
FRCNN-R101	FRCNN	81.8	25.7	44.5	18.9	29.7
Div	FRCNN	82.1	34.2	53.0	24.2	37.1
DivAlign	FRCNN	80.1	38.9	57.4	33.2	43.2
GDD	FRCNN	-	40.8	54.2	29.7	41.6
FGT	FRCNN	-	40.5	56.6	30.0	42.4
SE-COT	FRCNN	82.9	40.2	57.5	34.8	44.1
DiffDet	FRCNN	**86.2**	42.1	**60.9**	36.3	46.4
GS-Mix	FRCNN	83.4	**44.9**	60.8	**37.1**	**47.6**
YOLOv8s	YOLOv8	82.0	33.5	51.0	20.2	34.9
GS-Mix	YOLOv8	**83.0**	**44.9**	**58.1**	**32.9**	**45.3**

Table 3: Results of RT-DETR with GS-Mix on the normal-to-adverse weather task.

Methods	Detector	DS	NC	DR	NR	DF	mPC
RT-DETR-l	RT-DETR	56.5	43.0	35.6	20.8	40.7	35.0
GS-Mix	RT-DETR	**57.3**	**44.6**	**39.5**	**26.5**	**42.3**	**38.2**
RT-DETR-x	RT-DETR	56.3	44.4	36.8	22.0	41.2	36.1
GS-Mix	RT-DETR	**59.7**	**46.4**	**43.4**	**30.2**	**43.5**	**40.9**

Table 4: Results of medical domain generalization. DA uses unlabeled target data; S-DG uses source domain data only.

Methods	Detector	Train Type	Source (1000x)	Source (400x)	1000x →1000x	400x →400x
FRCNN	FRCNN	-	66.8	61.1	17.1	26.7
EMD	FRCNN	DA	-	-	37.5	33.8
YOLOv8s	YOLOv8s	-	69.1	61.5	35.1	30.1
GS-Mix	YOLOv8s	S-DG	**75.6**	**64.0**	**49.8**	**43.3**

Table 5: Results of the underwater domain generalization task. To provide a clearer comparison, we report the overall parameters (Params in M) and FLOPs (G) for the respective detectors.

Methods	Detector	Ho	Ec	Sc	St	mAP	Params (M)↓	FLOPs (G)↓
YOLOv8s	YOLOv8s	37.2	77.3	57.7	40.6	53.2	**11.2**	28.8
Hy-UOD	MFDA-Net	44.0	80.7	60.9	40.4	56.5	12.5	**28.1**
GS-Mix	YOLOv8s	**56.0**	**82.3**	**71.8**	**63.4**	**68.4**	**11.2**	28.8

On YOLOv8s, GS-Mix achieved an impressive 37.1% mPC. This result unlocks the potential of one-stage lightweight detectors, positioning them as compelling alternatives to Faster R-CNN due to their superior parameter efficiency and inference speed. This is because our GS-Mix operates exclusively at the pixel-level, overcoming the limitations of existing methods that rely on large-parameter backbones or two stage architectures. Additionally, compared to feature-level and frequency-domain augmentations, our approach delivers superior performance with lower implementation complexity.

Real-to-cartoon. As shown in Tab. 2, GS-Mix with Faster R-CNN achieves the best validation mAP on Clipart1K (44.9%) and Comic2k (37.1%), reaching 47.6% mPC with a 17.9% improvement over the baseline. It also achieves 45.3% mPC on lightweight YOLOv8s, demonstrating generalization across large style gaps. These results suggest that pixel-level structure-texture mixing reduces reliance on texture and style cues while promoting shape-based recognition.

Medical domain generalization. We implemented this task using the exact YOLOv8s configuration from the normal-to-adverse weather experiments, updating only dataset paths and class names. As shown in Tab. 4, GS-Mix improves upon the baseline detector by 14.7% mAP and 13.2% mAP at 1000x and 400x magnifications, respectively. Crucially, these enhancements are realized within a single domain generalization paradigm. In contrast to EMD (Sultani et al., 2022), which necessitates unlabeled target domain images for domain adaptation, GS-Mix demonstrates superior scalability and ease of deployment for downstream tasks.

Underwater domain generalization. As detailed in Tab. 5, GS-Mix achieves substantial performance gains over the SOTA method Hy-UOD (Luo et al., 2025) in challenging scenarios featuring low-contrast objects and cluttered scenes, all without requiring detector modifications. These gains stem from GS-Mix's noise and texture perturbations, which generate vicinal samples that direct the detector to extract object-relevant features in noisy, blurry scenes while suppressing domain-specific factors.

In addition, as illustrated in Tab. 3, to further verify the generalizability of GS-Mix across a broader range of detector architectures, we conducted experiments on RT-DETR (Zhao et al., 2024). The experimental results further confirm that even a purely pixel-level implementation of our method can significantly improve the generalization performance of the detector.

4.3 Ablation and Analysis

Training Cost Analysis. As detailed in Tab. 6, to better evaluate the efficiency of the proposed GS-Mix, we conducted experiments and analysis on training costs on a unified benchmark. The results indicate that the image space operations of GS-Mix incur significantly lower computational complexity and latency (45.1ms total; broken down as GCM: 4.9ms and STM: 40.2ms) compared to the frequency-domain method PhysAug (112.6ms). While this latency marginally exceeds that



Figure 4: Synthetic samples produced by the proposed STM. Despite the visually chaotic appearance, this effectively compels the model to mitigate its dependency on texture and focus on structures.

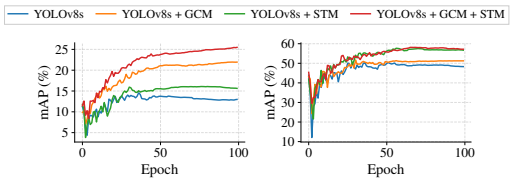

Figure 5: Performance comparison curves on Night Rainy (*Left*) and Watercolor2k (*Right*). The proposed component robustly improves training performance without abnormal fluctuations.

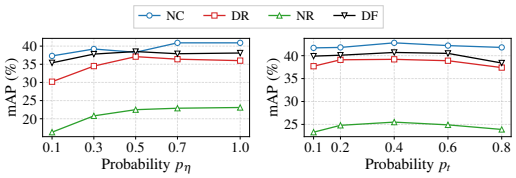

Figure 6: Hyperparameter analysis of GCM (*Left*) and STM (*Right*) mixing probabilities on the normal-to-adverse weather benchmark. p_t is tested with $p_\eta = 0.5$.

Table 6: Training cost comparison. Latency is measured on a single Intel Xeon Silver 4310 CPU core; ‡ denotes the worker count required to match GPU throughput at 5.3 ms/image.

Methods	Settings	Component	Latency (ms/img)	Workers to Match GPU‡	Workers = 8
Baseline	-	-	-	-	189 FPS
PhysAug	Fourier	FFT + IFFT	112.6	22	71 FPS
GS-Mix	Pixel	GCM + STM	45.1	9	182 FPS

Table 7: Ablation results on the proposed modules. Results cover four adverse weather domains (NC, DR, NR, and DF) and three cartoon domains (Cli., Wat., and Com.).

GCM	STM	NC	DR	NR	DF	Cli.	Wat.	Com.
-	-	41.7	27.9	14.6	35.3	33.5	51.0	20.2
✓	-	38.3	37.1	22.5	38.5	37.2	51.0	23.0
-	✓	**42.9**	32.2	16.1	35.3	42.0	57.4	**33.2**
✓	✓	42.8	**39.2**	**25.5**	**40.7**	**44.9**	**58.1**	32.9

of standard augmentations, the computational overhead is effectively masked by the GPU training duration when utilizing an asynchronous data loader with 9 workers.

Component analysis. We examined the GCM and STM components within GS-Mix under the basic settings of S-DGOD, as summarized in Tab. 7. Notably, GCM and STM deliver substantial improvements against noise perturbations and texture shifts, respectively, while their combination results in optimal performance. This demonstrates that GS-Mix effectively enables the detector to achieve OC-VRM through virtual samples generated by pixel-level intra-domain mixing. Moreover, the training performance curves presented in Fig. 5 corroborate the stability of these improvements.

Mixing probability analysis. We analyzed the probability thresholds for GCM and STM on the normal-to-adverse weather task, with results shown in Fig. 6. Observations indicate that conventional empirical configurations suffice to deliver substantial improvements.

5 Conclusion

We analyze the influencing factors and mechanisms of domain shift in S-DGOD and propose the OC-VRM distribution, which aims to synthesize samples that cover diverse and strong domain shift perturbations while preserving structure. Based on this, we construct GS-Mix, a simple, effective, and low-cost pixel-level mixing framework to instantiate this distribution. With only basic pixel-level intra-domain mixing, GS-Mix enables the detector to outperform state-of-the-art methods on S-DGOD benchmarks. Notably, it enables lightweight detectors to rival large-parameter models. In light of these findings, we envision this work serving as both a high-efficiency paradigm and a robust benchmark for future research in domain generalization.

AI USE STATEMENT

In this work, we used generative AI tools to aid in polishing the manuscript text and improving clarity, grammar, and readability. We also used AI tools for minor code editing and debugging assistance during implementation. All AI-assisted modifications were manually reviewed and verified by the authors. We take responsibility for the final content of this work.

REFERENCES

Meng Cao and Songcan Chen. Mixup-induced domain extrapolation for domain generalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10):11168–11176, Mar. 2024.

Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. In *Advances in Neural Information Processing Systems*, volume 13, 2000.

Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Meilin Chen, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Yunfeng Yan, Donglian Qi, Yueting Zhuang, Di Xie, and Shiliang Pu. Learning domain adaptive object detection with probabilistic teacher. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3040–3055, 17–23 Jul 2022.

Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3339–3348, June 2018.

Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. *CoRR*, abs/1805.09501, 2018.

Muhammad Sohail Danish, Muhammad Haris Khan, Muhammad Akhtar Munir, M. Saquib Sarfraz, and Mohsen Ali. Improving single domain-generalized object detection: A focus on diversification and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17732–17742, June 2024.

Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2918–2928, June 2021.

Boyong He, Yuxiang Ji, Zhuoyue Tan, and Liaoni Wu. Boosting domain generalized and adaptive detection with diffusion models: Fitness, generalization, and transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1912–1923, October 2025a.

Boyong He, Yuxiang Ji, Qianwen Ye, Zhuoyue Tan, and Liaoni Wu. Generalized diffusion detector: Mining robust features from diffusion models for domain-generalized detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9921–9932, June 2025b.

Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

Narges Honarvar Nazari and Adriana Kovashka. Domain generalization using shape representation. In *Computer Vision – ECCV 2020 Workshops*, pp. 666–670, Cham, 2020. ISBN 978-3-030-66415-2.

Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 749–757, March 2020.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Han Jiang, Wenfei Yang, Tianzhu Zhang, and Yongdong Zhang. Diffusion-based source-biased model for single domain generalized object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1548–1557, October 2025.

Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12456–12465, June 2019.

Jin-Seop Lee, Noo-ri Kim, and Jee-Hyong Lee. Domclp: Domain-wise contrastive learning with prototype mixup for unsupervised domain generalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(17):18119–18127, Apr. 2025.

Wooju Lee, Dasol Hong, Hyungtae Lim, and Hyun Myung. Object-aware domain generalization for object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4):2947–2955, Mar. 2024.

Wei Li, Boyu Zhao, Mengmeng Zhang, Yunhao Gao, and Junjie Wang. Single-source domain defect-aware adaptation and style-modulated generalization network for multispectral image segmentation. *IEEE Transactions on Cybernetics*, 56(3):1440–1451, 2026. doi: 10.1109/TCYB.2025.3624748.

Yajing Liu, Shijun Zhou, Xiyao Liu, Chunhui Hao, Baojie Fan, and Jiandong Tian. Unbiased faster r-cnn for single-source domain generalized object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28838–28847, June 2024.

Linxuan Luo, Pan Mu, and Cong Bai. Physics-coupled frequency dynamic adaptation network for domain generalized underwater object detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2284–2293, 2025. ISBN 9798400720352.

Massimiliano Mancini, Zeynep Akata, Elisa Ricci, and Barbara Caputo. Towards recognizing unseen categories in unseen domains. In *Computer Vision – ECCV 2020*, pp. 466–483, 2020.

Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *CoRR*, abs/1907.07484, 2019.

Poojan Oza, Vishwanath A. Sindagi, Vibashan VS, and Vishal M. Patel. Unsupervised domain adaptation of object detectors: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4018–4040, 2024. doi: 10.1109/TPAMI.2022.3217046.

Bin Pu, Xingguo Lv, Jiewen Yang, He Guannan, Xingbo Dong, Yiqun Lin, Li Shengli, Tan Ying, Liu Fei, Ming Chen, Zhe Jin, Kenli Li, and Xiaomeng Li. Unsupervised domain adaptation for anatomical structure detection in ultrasound images. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 41204–41220, 21–27 Jul 2024.

Waqas Sultani, Wajahat Nawaz, Syed Javed, Muhammad Sohail Danish, Asma Saadia, and Mohsen Ali. Towards low-cost and efficient malaria detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20655–20664, 2022. doi: 10.1109/CVPR52688.2022.02003.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30, 2017. doi: 10.1109/IROS.2017.8202133.

Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. Clip the gap: A single domain generalization approach for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3219–3229, June 2023.

Thanh Vu, Baochen Sun, Bodi Yuan, Alex Ngai, Yueqi Li, and Jan-Michael Frahm. Supervision interpolation via lossmix: Generalizing mixup for object detection and beyond. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5280–5288, Mar. 2024.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8052–8072, 2023. doi: 10.1109/TKDE.2022.3178128.

Zhitao Wen, Jinhai Liu, Huaguang Zhang, and Fengyuan Zuo. Exploring fine-grained visual-text feature alignment with prompt tuning for domain-adaptive object detection. *IEEE Transactions on Cybernetics*, 55(7):3220–3233, 2025. doi: 10.1109/TCYB.2025.3567126.

Aming Wu and Cheng Deng. Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 847–856, June 2022.

Aming Wu and Cheng Deng. Percept, memory, and imagine: World feature simulating for open-domain unknown object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4682–4691, June 2025.

Fan Wu, Jinling Gao, Lanqing Hong, Xinbing Wang, Chenghu Zhou, and Nanyang Ye. G-nas: Generalizable neural architecture search for single domain generalization object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5958–5966, Mar. 2024.

Xiaoran Xu, Jiangang Yang, Wenhui Shi, Siyuan Ding, Luqing Luo, and Jian Liu. Physaug: A physical-guided and frequency-based data augmentation for single-domain generalized object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(20):21815–21823, Apr. 2025.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: recaptioning, planning, and generating with multimodal llms. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Huaxiu Yao, Yiping Wang, Linjun Zhang, James Y Zou, and Chelsea Finn. C-mixup: Improving generalization in regression. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3361–3376. Curran Associates, Inc., 2022.

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 (ICCV)*, pp. 6023–6032, October 2019.

Seokju Yun, Seunghye Chae, Dongheon Lee, and Youngmin Ro. Soma: Singular value decomposed minor components adaptation for domain generalizable representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25602–25612, June 2025.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, October 2023.

Zihao Zhang, Aming Wu, and Yahong Han. Style evolving along chain-of-thought for unknown-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14225–14234, June 2025.

Hanqing Zhao, Dianmo Sheng, Jianmin Bao, Dongdong Chen, Dong Chen, Fang Wen, Lu Yuan, Ce Liu, Wenbo Zhou, Qi Chu, Weiming Zhang, and Nenghai Yu. X-paste: Revisiting scalable copy-paste for instance segmentation using CLIP and StableDiffusion. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 42098–42109, 23–29 Jul 2023.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16965–16974, June 2024.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Mixstyle neural networks for domain generalization and adaptation. *International Journal of Computer Vision*, 132(3):822–836, 2024.

A APPENDIX

In this supplementary material, we first provide implementation details, including detector configurations, training settings, and module parameters (Sec. A.1). We then present additional ablation studies on texture augmentation, edge thresholds, and mixing probabilities, together with comparisons against other augmentation methods (Sec. A.2). Next, we provide additional qualitative results on the normal-to-adverse weather, real-to-cartoon, underwater, and medical domain generalization tasks (Sec. A.3). Finally, we present the pseudo-code of GS-Mix, illustrating its integration into the YOLOv8s data augmentation pipeline (Sec. A.4).

A.1 IMPLEMENTATION DETAILS

We conducted experiments on the commonly used Faster R-CNN and YOLOv8 detectors. For Faster R-CNN, the backbone network is ResNet-101 pre-trained on ImageNet. Following the setup of previous work (Lee et al., 2024; Wu & Deng, 2022), we configure the SGD optimizer with a momentum of 0.9 and a weight decay of $1e^{-4}$, utilizing a batch size of 2 and a learning rate of $1e^{-3}$. For YOLOv8, we use the Small version as the baseline detector, following the official release configuration and using its pre-trained weights for initialization. We employ the SGD optimizer with a momentum of 0.937, an initial learning rate of $1e^{-2}$, and a batch size of 16. We use the standard mean average precision (mAP) at an IoU threshold of 0.5 as the evaluation metric, and report the mean performance (mPC) metric, which is the average of results across all target domains. The mixup parameter α is set to 32.0, and the probabilities p_η and p_t are set to 0.5 and 0.4, respectively. All experiments use consistent GCM and STM settings (including internal parameters).

In our implementation, for the lightweight detector baseline, we utilize the official Ultralytics YOLOv8s configuration. Its architecture features a CSP-based backbone, a PAFPN neck, and a decoupled anchor-free head. Its default data augmentation pipeline includes HSV augmentation (H: 0.015, S: 0.7, V: 0.4), image translation ($p = 0.1$), image scaling ($p = 0.5$), horizontal flip ($p = 0.5$), and mosaic augmentation ($p = 1.0$). In addition, we provide the specific parameters for STM in Tab. 8 to facilitate reproducibility.

A.2 MORE ABLATION EXPERIMENTS

Texture augmentation analysis. Tab. 9 substantiates the efficacy of our selected texture augmentation operators. All three operators consistently improve generalization on representative target domains (Night Rainy and Watercolor2k) across both tasks, albeit to different degrees. This indicates that mild texture augmentation is effective in S-DGOD for resisting potential target texture transformations, such as blur and color shift, thereby effectively covering the important vicinity space relevant to detection.

Edge-threshold analysis. Within the STM module, the Canny operator's gradient threshold dictates the density of edge extraction. To mitigate overfitting to specific edge sensitivities, we investigated both fixed empirical settings and a randomized interval sampling strategy. As illustrated in Fig. 7, while fixed empirical values yield consistent performance enhancements, randomized perturbations unlock superior generalization capabilities by introducing beneficial variance.

Parameter Experiments Analysis. We also report the experimental results for the real-to-cartoon task across varying mixing probabilities in Fig. 8. Notably, superior performance is attained by employing the identical probability configuration used in the normal-to-adverse weather task ($p_\eta =$

Table 8: Configuration of the STM module, including edge extraction and texture augmentation.

Modules	Operator	Param. Type	Configuration	Modules	Operator	Param. Type	Configuration
Edge Extra.	Canny	Hysteresis Thresh.	$L\in[50,150]$, $H\in[200,350]$	Texture Aug.	HSV Jitter	Saturation Gain	$\mathcal{U}(1.5, 2.5)$
	Dilation	Struct. Element	Ellipse, Iter=1		Quantization	Bit Depth	Choice $\in \{4, 5, 6\}$
	Gaussian Blur	Kernel Size	Fixed 5×5		Median Blur	Kernel Size	Choice $\in \{3, 5, 7\}$

Table 9: Ablation study of texture augmentation operators.

HSV jitter	Color Quat.	Median Blur	Night Rainy	Watercolor
-	-	-	24.2	57.1
✓	-	-	24.3	57.5
✓	✓	-	24.7	57.8
✓	✓	✓	**25.5**	**58.1**

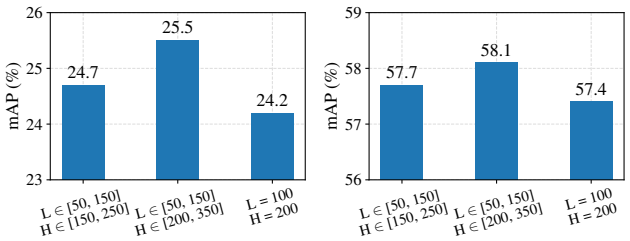

Figure 7: Edge threshold analysis on Night Rainy (*Left*) and Watercolor2k (*Right*). L: low threshold (gradients below are directly suppressed); H: high threshold (gradients above are directly preserved).

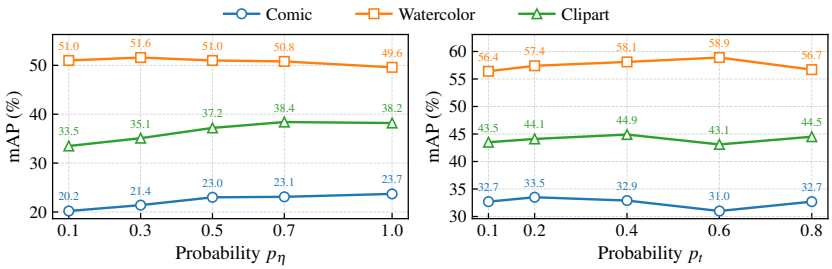

Figure 8: Hyperparameter sensitivity analysis of GCM (*Left*) and STM (*Right*) mixing probabilities on the real-to-cartoon benchmark. p_t is tested with $p_\eta = 0.5$.

0.5, $p_t = 0.4$). This robustness obviates the need for the laborious probability hyperparameter tuning typically required in conventional data augmentation pipelines.

Comparison with feature-level mixing. Tab. 10 details the comparative analysis against feature-level mixing approaches under YOLOv8s baseline. The results indicate that although such methods demonstrate efficacy in image classification, they yield negligible benefits in dense object detection scenarios. Conversely, our pixel-level mixing strategy delivers a substantial 6.6% mPC gain.

Comparison with other pixel-level augmentation. The Tab. 11 details our comparative analysis of mainstream pixel-level augmentations (CutMix (Yun et al., 2019), Copy-Paste (Ghiasi et al., 2021), and AugMix (Hendrycks et al., 2019)). Within our OC-VRM framework, CutMix and Copy-Paste can be viewed as constructing vicinal samples with spatial transformations, while AugMix mainly constructs vicinal samples with noise perturbations. Their weaker performance than GS-Mix on the current task suggests that, although they do not explicitly destroy structure, they also fail to introduce sufficiently strong perturbations, which limits the effective expansion of the vicinal sample space.

Comparison with other data augmentations on lightweight detectors. In Tab. 12 and Tab. 13, we reproduced and compared the performance of existing data augmentation-based S-DGOD methods on the one-stage detector YOLOv8s. We excluded the region-level feature consistency modules in DivAlign (Danish et al., 2024) and UFR (Liu et al., 2024), as they require ROI features from two-stage detectors. The results indicate that these methods yield limited generalization gains due to their ambiguous perturbation targets. Specifically, existing pixel-level vicinal sample synthesis methods mainly reduce structural damage by weakening augmentation strength, and therefore still remain within the conventional VRM setting rather than exploring an object-centric formulation.

Table 10: Comparison with feature-level mixing methods.

Methods	DS	NC	DR	NR	DF	mPC
AdaIN (Huang & Belongie, 2017)	55.8	40.9	31.5	16.6	33.1	30.5
MixStyle (Zhou et al., 2024)	55.4	41.8	31.4	17.3	35.9	31.6
GS-Mix	**56.1**	**42.8**	**39.2**	**25.5**	**40.7**	**37.1**

Table 11: Comparison with pixel-level augmentation methods on the normal-to-adverse weather task.

Methods	DS	NC	DR	NR	DF	mPC
CutMix (Yun et al., 2019)	56.0	42.3	30.0	14.6	35.6	30.6
Copy-Paste (Ghiasi et al., 2021)	**56.4**	41.5	27.6	12.3	34.5	29.0
AugMix (Hendrycks et al., 2019)	**56.4**	42.1	30.7	15.8	36.4	31.3
GS-Mix	56.1	**42.8**	**39.2**	**25.5**	**40.7**	**37.1**

A.3 Additional Visualization Results

We present further qualitative results in Fig. 9, Fig. 11, and Fig. 10 to verify the efficacy of GS-Mix in the S-DGOD task. As is evident from the visualizations, our method yields substantial performance gains for the detector under identical IoU thresholds.

Table 12: Comparison with data augmentation-based S-DGOD methods on the real-to-cartoon task.

Methods	VOC	Cli.	Wat.	Com.	mPC
UFR (Liu et al., 2024)	81.9	37.9	52.9	23.1	38.0
Div (Danish et al., 2024)	80.9	39.1	53.2	26.1	39.4
PhysAug (Xu et al., 2025)	82.5	38.0	55.0	25.1	39.4
GS-Mix	**83.0**	**44.9**	**58.1**	**32.9**	**45.3**

Table 13: Comparison with data augmentation-based S-DGOD methods on the normal-to-adverse weather task.

Methods	DS	NC	DR	NR	DF	mPC
UFR (Liu et al., 2024)	54.9	41.0	32.8	18.3	35.5	31.9
Div (Danish et al., 2024)	56.0	40.8	35.1	21.9	34.7	33.1
PhysAug (Xu et al., 2025)	55.3	**42.9**	37.9	20.8	37.0	34.7
GS-Mix	**56.1**	42.8	**39.2**	**25.5**	**40.7**	**37.1**

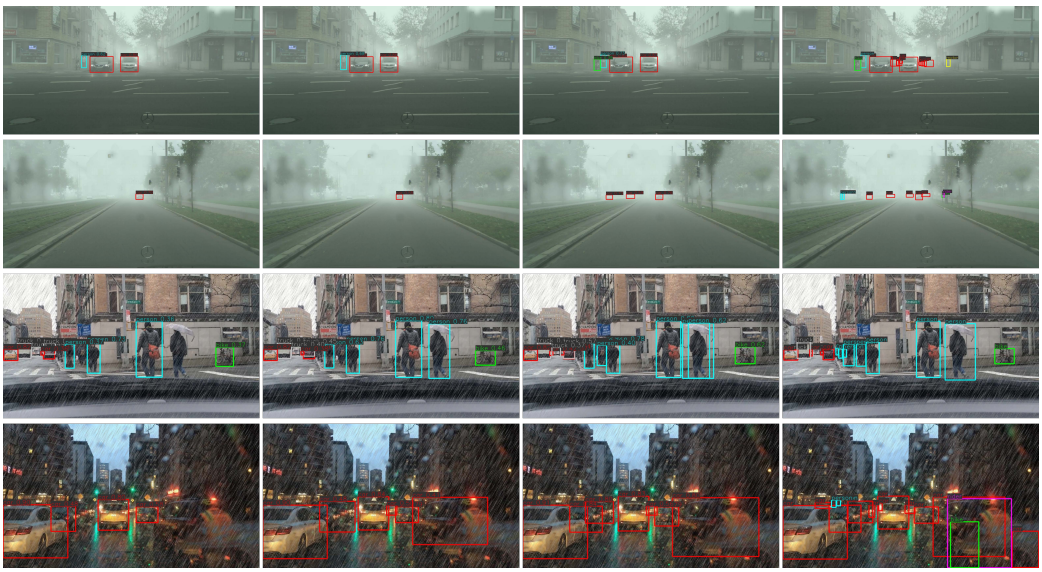

Figure 9: Qualitative evaluation on the normal-to-adverse weather task. Columns display (L-R): YOLOv8s baseline, MixStyle (Zhou et al., 2024), GS-Mix (Ours), and Ground Truth. All visualizations utilize an IoU threshold of 0.35. Our method demonstrates robust generalization across diverse adverse weather conditions, exhibiting superior recognition and localization capabilities even under severe noise perturbations like dense fog and rain streaks.

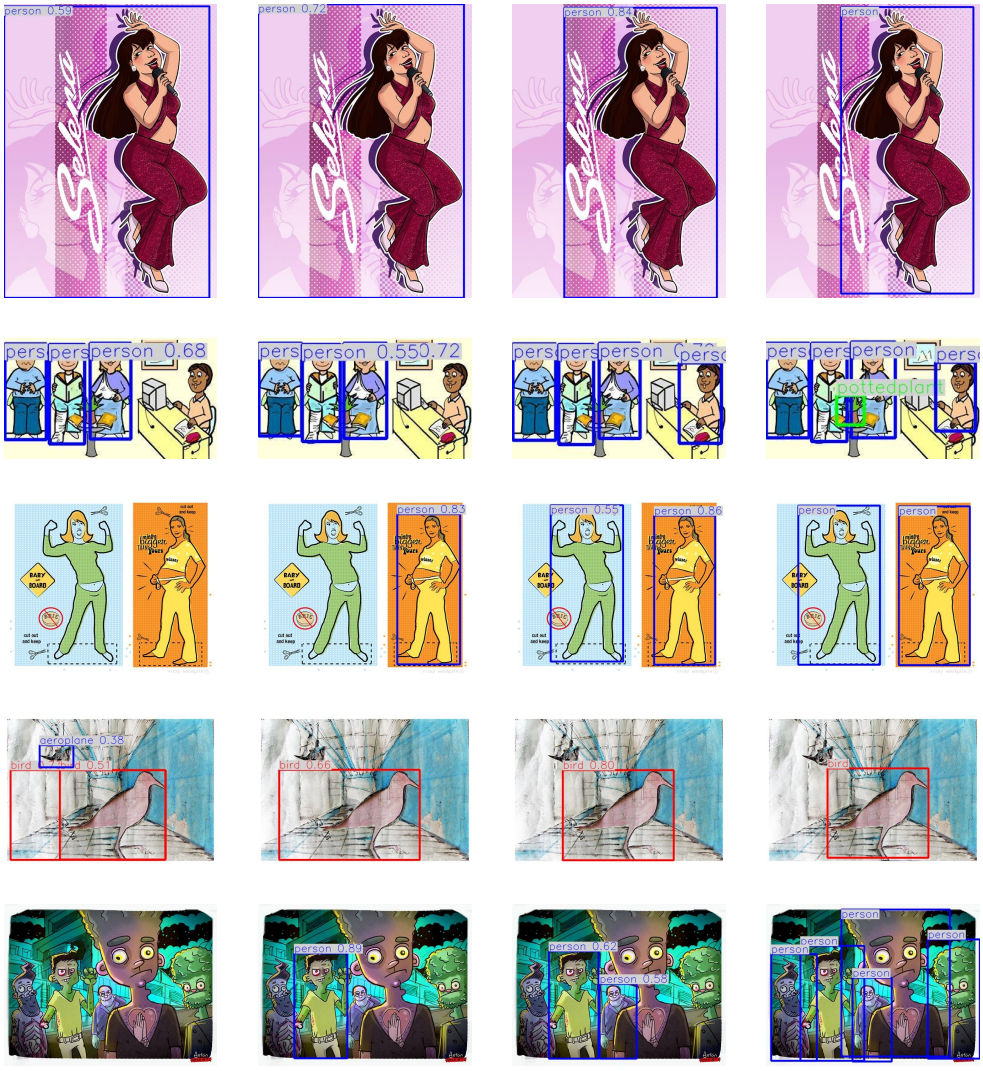

Figure 10: Qualitative results on the real-to-cartoon task. Columns display (L-R): YOLOv8s baseline, MixStyle, GS-Mix (Ours), and Ground Truth. All visualizations utilize an IoU threshold of 0.35. Despite the severe texture perturbations inherent in cartoons relative to the VOC training data, our GS-Mix yields superior bounding box regression performance.

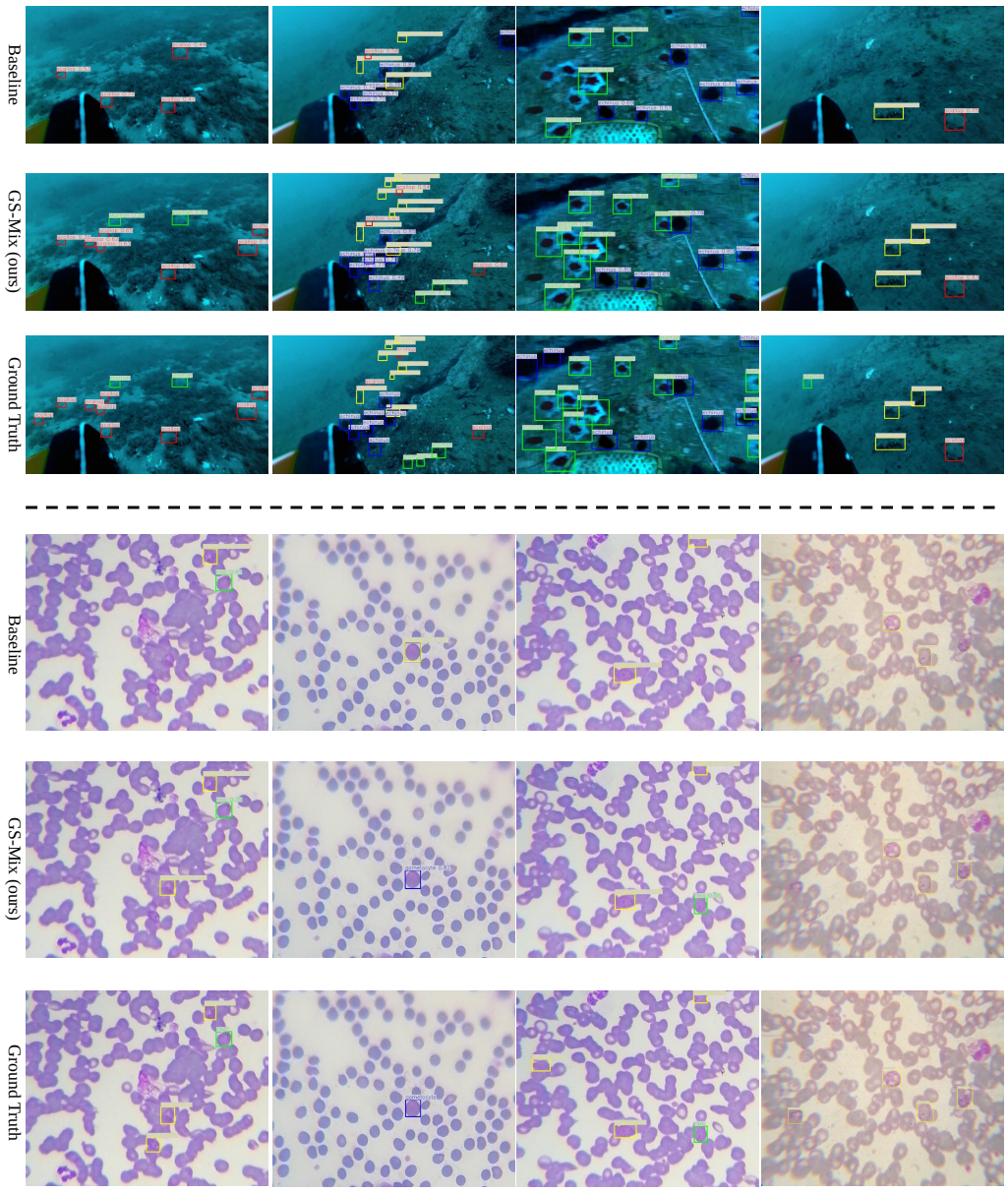

Figure 11: Qualitative comparison on the underwater (top) and medical (bottom) domain generalization benchmarks. Visual samples compare the YOLOv8s baseline, GS-Mix (Ours), and Ground Truth. In challenging underwater environments suffering from low contrast and indistinct object contours, our approach significantly boosts the baseline's performance. Furthermore, GS-Mix achieves better recognition of target objects even under low-cost microscopes with blurry imaging.

Algorithm 1 GS-Mix Framework

Input: Dataset $\mathcal{S}$, Mosaic size $L = 4$, Bernoulli param π, STM prob p_t, GCM prob p_η
Output: Augmented image $\tilde{x}_{mos}$, Labels $\tilde{y}_{mos}$
Initialize empty canvas x_{mos}, label set $y_{mos} \leftarrow \emptyset$
{Path Selection: 0 for STM path, 1 for GCM path}
Sample selector $\gamma \sim \text{Bernoulli}(\pi)$
for $i = 1$ **to** L **do**
  Sample image-label pair (x_i, y_i) from $\mathcal{S}$
  {**Stage 1: STM (Structural Texture Mixing)**}
  // Activate STM only if $\gamma = 0$ (STM path) and random prob met
  if $\gamma = 0$ **and** $\text{Random}(0, 1) < p_t$ **then**
    Sample texture source (x_j, y_j) from $\mathcal{S}$
    // Structure Extraction
    $\mathbf{m}(x_i) \leftarrow \text{Canny}(x_i) \rightarrow \text{Dilate} \rightarrow \text{GaussianSmooth}$
    // Texture Augmentation
    $x_j^a \leftarrow \text{HSV}(x_j) \rightarrow \text{Quantization} \rightarrow \text{MedianBlur}$
    // Structure-Preserving Mixing
    $\tilde{x}_i^t \leftarrow x_i \odot \mathbf{m}(x_i) + x_j^a \odot (1 - \mathbf{m}(x_i))$
    $y_i \leftarrow y_i \cup y_j$
    $x_i \leftarrow \tilde{x}_i^t$
  end if
  Calculate coordinates for patch i in Mosaic
  Place x_i into x_{mos} and update labels $y_{mos} \leftarrow y_{mos} \cup y_i$
end for
{**Stage 2: GCM (Global Context Mixup)**}
// Activate GCM only if $\gamma = 1$ (GCM path) and random prob met
if $\gamma = 1$ **and** $\text{Random}(0, 1) < p_\eta$ **then**
  Sample mixing image (x_{mix}, y_{mix}) from $\mathcal{S}$
  Sample $\lambda \sim \text{Beta}(\alpha, \alpha)$
  $x_{mos} \leftarrow \lambda x_{mos} + (1 - \lambda) x_{mix}$
  $y_{mos} \leftarrow y_{mos} \cup y_{mix}$
end if
return x_{mos}, y_{mos}

A.4 Pseudo-Code of GS-Mix

We present the implementation pseudo-code for GS-Mix within YOLOv8s in Algorithm 1. This serves to illustrate its seamless integration into existing data augmentation frameworks characterized by geometric transformations (e.g., Mosaic), thereby demonstrating its deployment flexibility.