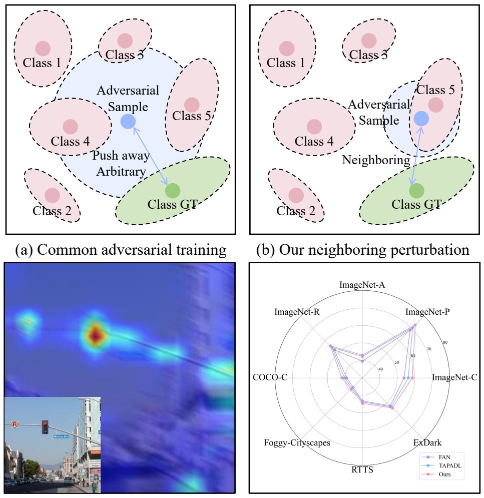
Gain from Neighbors: Boosting Model Robustness in the Wild via Adversarial Perturbations Toward Neighboring Classes

Zhou Yang[1] Mingtao Feng[1†] Tao Huang[1] Fangfang Wu[1] Weisheng Dong[1†]
Xin Li[2] Guangming Shi[1]
[1]Xidian University [2]University at Albany, State University of New York
yang_zhou@stu.xidian.edu.cn mintfeng@hnu.edu.cn {wsdong,gmshi}@mail.xidian.edu.cn

Abstract

*Recent approaches, such as data augmentation, adversarial training, and transfer learning, have shown potential in addressing the issue of performance degradation caused by distributional shifts. However, they typically demand careful design in terms of data or models and lack awareness of the impact of distributional shifts. In this paper, we observe that classification errors arising from distribution shifts tend to cluster near the true values, suggesting that misclassifications commonly occur in semantically similar, neighboring categories. Furthermore, robust advanced vision foundation models maintain larger inter-class distances while preserving semantic consistency, making them less vulnerable to such shifts. Building on these findings, we propose a new method called **GFN** (**G**ain **F**rom **N**eighbors), which uses gradient priors from neighboring classes to perturb input images and incorporates an inter-class distance-weighted loss to improve class separation. This approach encourages the model to learn more resilient features from data prone to errors, enhancing its robustness against shifts in diverse settings. In extensive experiments across various model architectures and benchmark datasets, GFN consistently demonstrated superior performance. For instance, compared to the current state-of-the-art TAPADL method, our approach achieved a higher corruption robustness of 41.4% on ImageNet-C (+2.3%), without requiring additional parameters and using only minimal data.*



(a) Common adversarial training (b) Our neighboring perturbation

(c) Class attention map (d) Performance

Figure 1. (a) and (b) Difference between common adversarial method and ours. (c) The attention map of our method on motion blur image. (d) Robust performance on several benchmark datasets. Adversarial samples generated by neighboring perturbations can effectively boost the model robustness trained by empirical risk minimization (ERM).

1. Introduction

Model training traditionally relies on empirical risk minimization (ERM), a process focused on enhancing a model's ability to generalize to unseen test data by minimizing loss on the training set. ERM [49] typically assumes that training and test data are independently and identically distributed (**IID**). However, this assumption is challenging to meet in many real-world applications. When faced with out-of-distribution (**OOD**) data—such as corrupted images, adversarial perturbations, or style variations—the model's performance often deteriorates sharply, as it has not been trained to handle these distribution shifts. Therefore, enhancing model robustness involves designing approaches to maintain optimal performance on the test set, even when the training and test data distributions diverge. This can be formulated as,

$$\begin{aligned} \boldsymbol{f}_{\boldsymbol{\theta}}^{*} &= arg\min_{\boldsymbol{f}_{\boldsymbol{\theta}}} \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}\in P_{te}}[\mathcal{L}(\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{X}),\boldsymbol{Y})] \\ & s.t. P_{te}(\boldsymbol{X},\boldsymbol{Y}) \neq P_{tr}(\boldsymbol{X'},\boldsymbol{Y'}), \end{aligned} \tag{1}$$

[†] Corresponding authors.

where $\boldsymbol{f_\theta}$ represents the model with parameters $\boldsymbol{\theta}$, $P_{te}(\cdot)$, $P_{tr}(\cdot)$ denoted as the test and training data distribution, respectively.

One line of research on robust learning involves designing sophisticated data augmentation strategies that utilize the existing training set to generate data with varied distributions, helping the model learn robust representations and thereby improving its ability to handle out-of-distribution samples. Traditional data augmentation methods such as Mixup [56] and CutMix [55] create new distributions by blending two images proportionally or by cutting and pasting parts of different images. In contrast, adversarial training [35] focuses on generating challenging examples by applying gradient-based perturbations to the original data. However, these methods lack clear insights into which specific types of data distribution perturbations most effectively enhance model robustness, thus limiting interpretability. Recently, some approaches involve modifying the entire transformer architecture or adding new modules at the token level. For example, FAN [60] analyzed vision transformers (ViT) using information bottleneck theory and identified the attention layer as a key factor influencing model robustness, leading to the development of a novel full-attention network. TAPADL [14] aimed to reduce the impact of noise in input images by incorporating a token average pooling layer to smooth the inputs. However, it is worth noting that these modifications may be non-standard, and certain operations may lack support for parallelism, resulting in a substantial increase in computation time, which can limit their viability as general-purpose solutions.

To address these issues, we must consider the following questions: 1) What effects does distribution shift have on model outputs, especially in classification tasks? 2) How can we generate training data that effectively induces similar effects? Based on Eq. (1) and the principle of Empirical Risk Minimization (ERM), if a model is trained on data from diverse distributions that closely approximate potential test distributions, its robustness will likely improve. 3) What strategies can leverage the generated data to enhance the model's robustness, particularly through refined loss functions? Accordingly, we designed a series of experiments to investigate these questions. From the experimental results and statistical analyses, we conclude that *most classification errors due to distribution shifts are primarily concentrated near the true values*, meaning that misclassifications tend to fall into categories semantically similar to the ground-truth labels. Additionally, we observed that *advanced foundation models exhibit greater inter-class distances compared to more vulnerable models*, contributing to improved robustness.

Building on these observations, we propose a gradient perturbation method that targets neighboring classes, generating valuable training data by applying these perturbations to the original input images. These gradient-based perturbations increase the likelihood of the data being misclassified into neighboring classes, effectively simulating the out-of-distribution (OOD) conditions the model may face during testing. Additionally, we introduce an inter-class distance-weighted loss to fine-tune the model parameters, enhancing the model's ability to make accurate predictions even with perturbed data. This approach also helps to create a more dispersed class distribution, improving the model's overall robustness. In summary, the main technical contributions are listed below.

- We investigated the impact of distribution shift data on model outputs through experiments and observed that erroneous predictions are often semantically similar categories. This provides valuable insights for the subsequent design of training data.
- We proposed a novel gradient perturbation adversarial training method that targets gradient attacks specifically at neighboring classes. These perturbed images, which are more likely to be misclassified, mimic the distribution of OOD data, and their inclusion in training can improve model robustness effectively.
- We designed a novel inter-class distance-weighted loss function, which assigns greater weight to classes that are more similar to the ground-truth class. This encourages the learned feature representations from different classes to become more dispersed, thereby enhancing the model's robustness.
- Extensive experimental results demonstrated that our approach more efficiently improves model accuracy on several benchmark out-of-distribution datasets compared to the state-of-the-art methods.

2. Related Work

2.1. Robust Learning

Robust learning aims to enhance the model's transferability and generalization capabilities, ensuring strong performance across datasets with varying distribution. The primary related works can be categorized into the following three types.

Data Augmentation: Data augmentation [31, 33, 40, 51] is a technique that leverages existing data to generate new samples with different distributions through manually designed transformations, aiming to improve the model's generalization performance. Conventional methods generate augmented images through transformations such as rotation, scaling, and translation, often combining them using various strategies. Mixup [56], and CutMix [55] combine new augmented images through linear combinations or cropping & pasting. APR [4] generated augmented images by swapping the amplitude and phase spectra of the image's frequency domain. Augmix [18] randomly performed different data augmentations on images and then mixed them to form the final augmented output. AutoAugment [7] first used reinforcement learning to find the optimal data augmentation strategy. Recently, the emergence of generative networks has led to the use of models to generate additional data for augmentation during training. DeepAugment [19] used a generative adversarial network (GAN), while Diffmix [22] and [12] used diffusion models to generate the augmented images.

Adversarial Training: Adversarial training originates from adversarial attacks on neural networks. Its main idea is to generate gradient-based adversarial perturbations for the target network, causing the perturbed input images to produce erroneous outputs. Adversarial training, in turn, involves using such perturbed images to make the model more robust. Buckman et al. [2] and Su et al. [46] defined the adversarial training as a min-max problem:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\max_{\boldsymbol{x'},\ ||\boldsymbol{x'}-\boldsymbol{x}||\leq\epsilon} \mathcal{L}_{ce}(\boldsymbol{\theta}, \boldsymbol{x'}, y) \right], \tag{2}$$

where $\boldsymbol{x'}$ is a perturbed image, $\boldsymbol{\theta}$ represents the model parameters, ϵ is used to control the range of perturbations. To ensure that the perturbed image is close to the original one while satisfying the aforementioned min-max conditions, $\boldsymbol{x'}$ is defined as follows:

$$\boldsymbol{x'} = \boldsymbol{x} + \eta * sign(\nabla_{\boldsymbol{x}} \mathcal{L}_{ce}(\boldsymbol{\theta}, \boldsymbol{x}, y)), \tag{3}$$

where η is the step size and ∇ represents the gradients. Many studies [1, 6, 9, 35, 39, 53, 59] have demonstrated that adversarial training effectively enhances the robustness of models against adversarial attacks.

Robust Model Architecture: Some methods explore the model robustness from the architecture itself. Bai et al. [1] first investigated the robustness between transformers and CNNs, demonstrating that ViTs have better robust performance. FAN [60] supposed that self-attention is a key component of robustness and proposed a new fully attentional network. RVT [36] proposed a robust vision transformer through a new position embedding strategy and a patch-wise augmentation method. TAPADL [14] proposed adding extra convolutional layers before the transformer to enhance the model's robustness against noisy images.

2.2. Real World Recognition

Recently, many studies [8, 11, 15, 16, 25, 28, 34, 44, 57, 58] have focused on recognition tasks in real-world scenarios characterized by adverse weather conditions, such as rain, fog, snow, and low-light environments. Researchers [27, 37, 45, 47] enhance the recognition performance of those corrupted images by combining image restoration with classification or detection tasks. Others [23, 52, 54] improve model robustness from the perspective of image features. For example, Ada-YOLO [29] proposed a DIP module to restore a clear image from a hazy one. Then, the degraded image object detection was treated as a multi-task learning [3] problem (restoration and detection) and jointly optimized. FeatEnhancer [16] proposed a feature enhancement network to improve the hierarchical latent representation of low-light images. Recently, some studies [21, 50] have leveraged the Contrastive Language–Image Pretraining (CLIP) [41] model to enhance the model's generalization performance through language guidance.

3. Method

3.1. Motivations

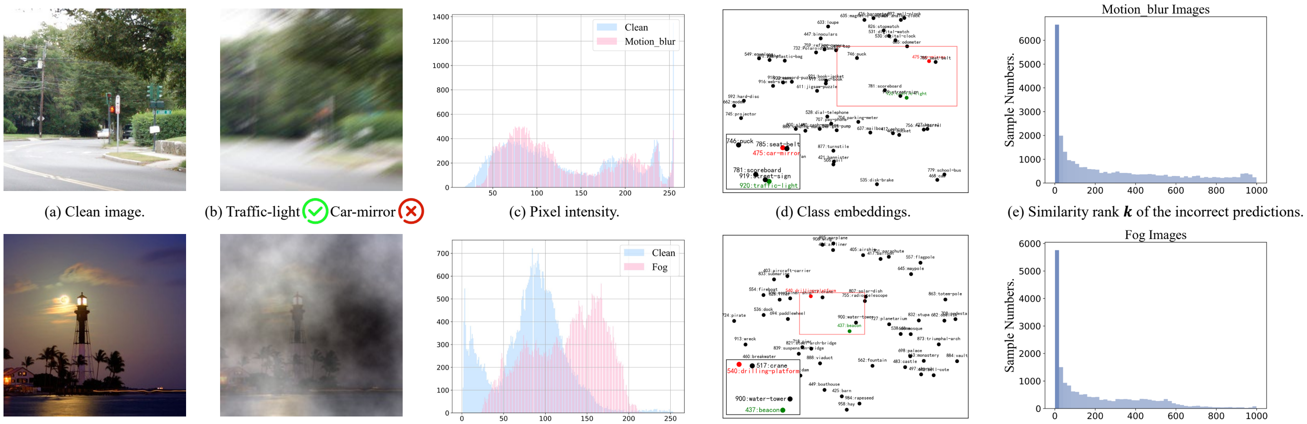
Numerous studies have highlighted that a model's performance can degrade significantly on out-of-distribution data. We aim to determine the properties of the erroneous outputs and these OOD data's underlying distributions. To investigate this, we conducted the following experiments.

Error Predictions Analysis: We first explore the impact of OOD data on model prediction. For a pre-trained ViT-Base model $\boldsymbol{f_\theta}$ with parameter $\boldsymbol{\theta}$, we have a clean image $\boldsymbol{x}$ with label c in the ImageNet test set and a corresponding paired degraded image $\tilde{\boldsymbol{x}}$ in the ImageNet-C dataset. We input these out-of-distribution degraded images into the model for inference and collect images correctly classified when using clean data but misclassified when using corrupted ones (i.e., wrong set $S = \{\tilde{\boldsymbol{x}} | \boldsymbol{f_\theta}(\tilde{\boldsymbol{x}}) = y_p \neq y_c, where\ \boldsymbol{f_\theta}(\boldsymbol{x}) = y_c\}$). This ensures that in S, the misclassifications are solely attributed to degradation effects rather than inherent model limitations on some clean data.

Figs. 2 (a) and 2 (b) show the clean images, the ground truth labels, and the corresponding wrong predictions of the degraded images. We observed that the misclassified category is semantically similar to its label (e.g., beacon and drilling platform), and we also visualized the class embeddings (mean / center feature of each class) in the feature space, as shown in Fig. 2 (d), which further supports this conclusion. Additionally, we conducted extensive statistical analyses across multiple datasets to rule out coincidence. For each class embedding $\boldsymbol{z_c} \in \mathbb{R}^{1\times d}$, we compute its similarity with others, resulting in a similarity matrix $\mathcal{M} \in \mathbb{R}^{C\times C}$, where $\mathcal{M}_{ij} =< \boldsymbol{z_i}, \boldsymbol{z_j} >, i, j \in \{1, \ldots, C\}$ represents the similarity between class i and j. Then for a image $\tilde{\boldsymbol{x}} \in S$, we can obtain the similarity rank index k between its wrong prediction class p and its label c through:

$$k = rank(\mathcal{M}_{cp}), \mathcal{M}_{cp} \in sorted(\mathcal{M}_{c1}, \ldots, \mathcal{M}_{cC}). \tag{4}$$

That is, the misclassification occurs within the k-th most similar class to the ground truth. For each sample in S, we computed the rank k, and the resulting histogram of sample



(a) Clean image. (b) Traffic-light ✓ Car-mirror ✗ (c) Pixel intensity. (d) Class embeddings. (e) Similarity rank ***k*** of the incorrect predictions.

(a) Clean image. (b) Beacon ✓ Drilling-platform ✗ (c) Pixel intensity. (d) Class embeddings. (e) Similarity rank ***k*** of the incorrect predictions.

Figure 2. Observations on wrong predictions generated by corrupted data. (a) The original high-quality / clean images. (b) The target label and the misclassification output of "motion-blue" and "fog" corrupted images. (c) The pixel intensity of clean and corrupted images. (d) The class embeddings of the **top-50 similar categories** to the ground-truth label. Green represents the ground truth, while red indicates the misclassified predictions. The number preceding the text corresponds to the class index in ImageNet-1k, ranging from 0 to 999. (e) A statistical analysis of sample numbers and similarity rank k of incorrect predictions. We conducted experiments on misclassifications caused by 15 types of corruptions in ImageNet-C [17] and plotted the results for motion blur and fog images. **Zoom in for better details.**

numbers for rank k is shown in Fig. 2 (e). More results can be found in the supplementary materials.

These figures show that most erroneous predictions are concentrated within the top k nearest classes, with many samples falling within similarity rank $k < 100$. This indicates that the predicted classes frequently exhibit a close semantic or feature similarity to the true class when errors arise due to the corrupted data in the wild.

Distribution Shift: We also computed the pixel value intensity of clean / high-quality images and corrupted ones. As shown in Fig. 2 (c). It can be observed that the overall distribution of pixel values remains largely unchanged. It exhibits a uniform shift, either increasing or decreasing. This suggests that the perturbations primarily affect the overall intensity of the pixel values without significantly altering the underlying structure of the images. In other words, the changes introduced by the perturbations are more related to global shifts in brightness or contrast rather than fundamentally modifying the spatial relationships or patterns in the image. This might explain why models misclassify these images into semantically similar categories, as the essential features of the image remain largely intact.

Inter-Class Similarity: To further explore the characteristics of more robust models, we visualized the inter-class similarity maps of $\mathcal{M}$ for both less robust ViT-Base model and the highly robust DINOv2 [38] model. DINOv2 pretrained on a large-scale dataset LVD-142M [38] in a self-supervised manner and has a clean accuracy of 84.6% and robust accuracy of 72.1% on ImageNet-C.

As shown in Fig. 3, we can conclude that, for the ViT-Base model, certain classes exhibit higher similarity to their neighboring classes. However, this pattern is less pronounced in the case of DINOv2. This indicates that misclassifications are more likely to occur between these similar classes in common ViTs, but the DINOv2 model has learned more distinct inter-class representations, reducing the likelihood of confusion between neighboring classes.

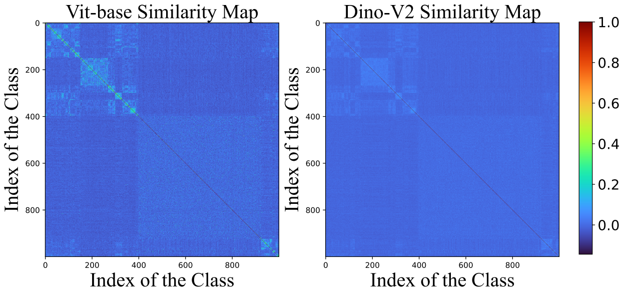


Figure 3. Visualization of inter-class cosine similarity for the ViT-Base and DINOv2 classification models.

In summary, the above observations suggest that errors are not entirely random but follow a pattern where the incorrect predictions are close to the true class. And robust models exhibit smaller inter-class similarities, corresponding to larger inter-class distances, making them less prone to misclassifications.

3.2. Learning from Misclassified Neighboring Data

The observations in Sec. 3.1 provide valuable guidance for designing more effective training strategies, such as for a

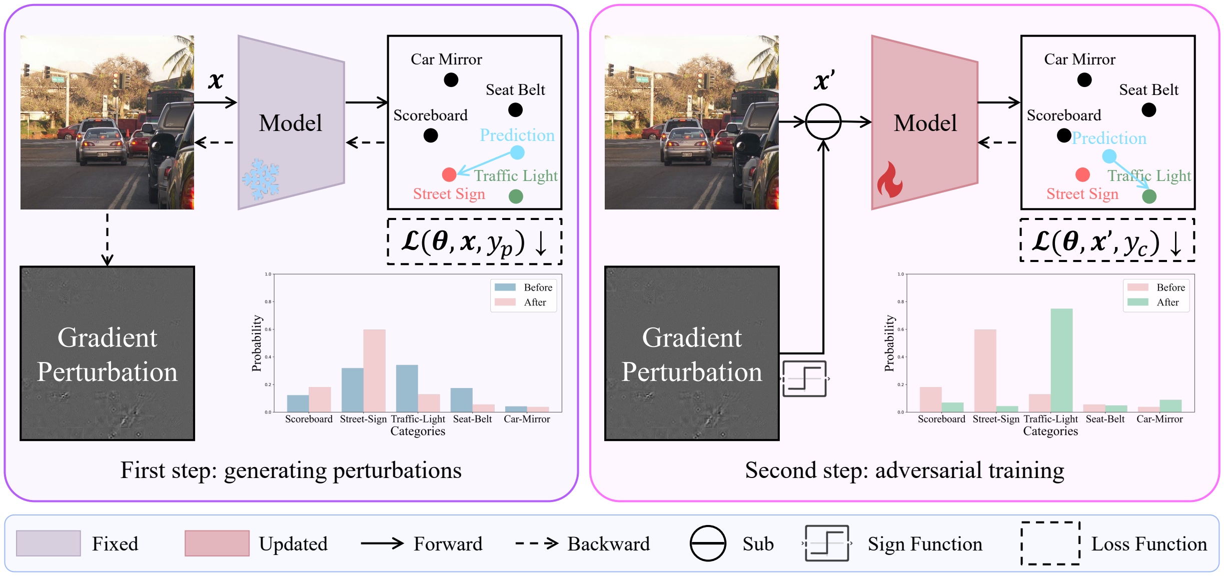


Figure 4. The overall architecture of our proposed method **GFN**. The blue point in the embedding space represents the predicted values, the red indicates the neighboring class, and the green represents the ground truth. The probability plot illustrates the change from $p(y|\boldsymbol{x})$ to $p(y|\boldsymbol{x}')$ and then to the adversarial trained $p(y|\boldsymbol{x}')$ as described in Sec. 3.2. **Best viewed in color. Zoom in for better details.**

given image $\boldsymbol{x}$, we can generate a perturbed version $\boldsymbol{x}'$ that is prone to be misclassified into neighboring classes. By targeting these near-neighbor misclassifications, we can better simulate the impact of distribution shifts and enhance the model's robustness. Therefore, given a pre-trained network $\boldsymbol{f_\theta}$ and a clean input image $\boldsymbol{x}$ with label y_c, our goal is to generate an image $\boldsymbol{x}'$ such that $\boldsymbol{f_\theta}(\boldsymbol{x}') = y_p$, where p is among the top k categories most similar to c. Here, we use the input gradients to achieve this effectively. We have,

$$\boldsymbol{x}' = \boldsymbol{x} - \eta * sign(\nabla_{\boldsymbol{x}} \mathcal{L}_{ce}(\boldsymbol{\theta}, \boldsymbol{x}, y_p)), \tag{5}$$

where η is the step size, $\mathcal{L}_{ce}$ is the cross entropy loss and $\nabla_{\boldsymbol{x}}$ represents the gradients of input $\boldsymbol{x}$. Since the negative gradient minimizes the loss, it can be expected that $\boldsymbol{x}'$ will bias the model toward predicting neighboring class p.

Subsequently, we optimize the model by constraining it to produce correct results (i.e., class c) on the perturbed data $\boldsymbol{x}'$, which can be formulated as,

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \ \mathcal{L}_{ce}(\boldsymbol{x}', y_c). \tag{6}$$

To summarize, we first generated gradient perturbed images using Eq. (5) and then applied Eq. (6) to ensure that these images produced correct outputs. As shown in Fig. 1 and Eq. (3), it is important to note that this differs from conventional adversarial training, which aims to push the output away from the true value arbitrarily. In contrast, our approach intentionally guides the output toward neighboring classes, making it more focused and effective. This training approach enables the model to learn the latent structure of potential out-of-distribution data efficiently, enhancing its robustness by improving its ability to generalize across varying data distributions.

3.3. Inter-Class Distance Weighted Loss

The neighbor sample adversarial training method described in Sec. 3.2 generates perturbation data to help the model distinguish between similar categories. Combining this with Sec. 3.1 and Fig. 3, another insight is to enhance model robustness by increasing the inter-class distance to further reduce category confusion.

Consequently, we propose to weight the loss function based on the inter-class distances. It penalizes misclassifications among closely related neighboring classes more heavily, thereby enhancing the model's ability to differentiate between these similar categories. So we designed a simple weighting term as,

$$\boldsymbol{w} = (w_1, \ldots, w_p, \ldots, w_C), w_p = \alpha^k, 0 < \alpha < 1, \tag{7}$$

where α is a hyperparameter, we set $\alpha < 1$, and k represents the similarity rank between the predicted class p and the true label c, as shown in Equation (4). The higher the similarity, the smaller the k value and, consequently, the larger the weight. For example, the weight for the 2nd similar class is $0.5^2 = 0.25$, while the weight for the 20th similar class is $0.5^{20} \approx 9e-7$. This weight penalizes more similar classes, effectively increasing inter-class separation and enhancing the model's robustness.

Method	# Param (M)	# FLOPS (G)	# T (ms)	ImageNet ↑	ImageNet-C ↓	ImageNet-P ↓	ImageNet-A ↑	ImageNet-R ↑
ViT-B [10]	86.4	16.8	4.8	76.3	52.2	32.2	31.7	50.7
DeiT-B [48]	86.4	16.8	4.8	78.4	47.4	31.8	27.6	45.3
Swin-B [30]	87.8	15.4	17.3	83.4	54.4	32.7	35.8	46.6
RVT-B (ViT-B) [36]	91.8	17.7	5.3	82.6	46.8	31.9	28.5	48.7
FAN-B-Hybrid [60]	50.4	11.7	20.8	83.9	46.1	31.3	39.6	52.7
RSPC (FAN) [13]	50.5	11.7	21.0	84.2	44.5	30.0	41.1	53.4
TAPADL (FAN) [14]	50.7	11.8	21.5	84.3	43.7	29.2	42.3	54.6
Ours (ViT-B)	86.4	16.8	4.8	79.9	41.4	27.1	42.0	55.9
Ours (FAN)	50.4	11.7	20.8	**86.4**	**39.8**	**26.3**	**46.4**	**57.7**

Table 1. Comparison with other methods on ImageNet and related robustness benchmark datasets. T represents the time taken for the forward propagation of each method over 100 iterations. The best results are indicated in **bold**, while the second ones are underlined.

3.4. Overall Framework and Training Process

The overall framework is illustrated in Fig. 4. Since training a model from scratch is time-consuming, we leverage the off-the-shelf pre-trained model for efficient fine-tuning. In the first step, for a given clean image $\boldsymbol{x}$ with label y_c, we first set it to be differentiable (with respect to gradients, i.e., `x.requires_grad=True`) and input it into the model. We randomly select a class p from the top-k most similar classes as the target, compute the cross-entropy loss, and then perform backpropagation based on Eq. (5) to obtain the gradient and perturbed image $\boldsymbol{x'}$. In this process, the model is only involved in backpropagation without updating its parameters. Secondly, We feed the perturbed image $\boldsymbol{x'}$ back into the model and fine-tune its parameters to correctly classify $\boldsymbol{x'}$ into true label c. This is done by minimizing inter-class distance-weighted classification loss,

$$\mathcal{L}_{ce}(y_c, \boldsymbol{f_\theta}(\boldsymbol{x'})) = -y_c * \boldsymbol{w} * \log(softmax(\boldsymbol{f_\theta}(\boldsymbol{x'}))). \quad (8)$$

This process can be iterated with different images and target classes, ensuring the model consistently classifies perturbed images back to their correct labels.

4. Experiments

4.1. Datasets and Implementation Details

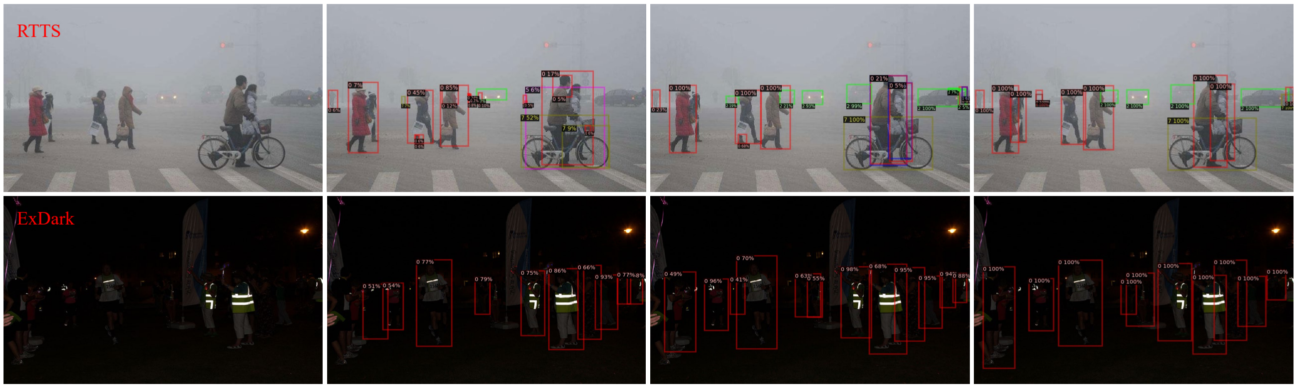
Datasets: In experiments, we followed the settings used in the previous model robustness methods. Specifically, in classification, we trained exclusively on high-quality / clean ImageNet images and tested them on out-of-distribution datasets such as ImageNet-C [17], ImageNet-P [17], ImageNet-R [19], and ImageNet-A [20]. The ImageNet-C benchmark dataset contains 15 types of common corruptions. ImageNet-P departs from Imagenet-C by generating perturbation sequences from each ImageNet validation image. The evaluation metrics used are the normalized mean Corruption Error (mCE) and mean Flip Rate (mFR). Lower values indicate better performance. ImageNet-R contains sketches, art, paintings, toys, and other style images, while ImageNet-A provides adversarial examples for testing model robustness on adversarial attacks; we reported the top-1 accuracy for these two datasets.

Implementation Details: Since the transformer architecture is widely used, our experiments are conducted based on vision-transformer (ViT) models. We employed SGD [42] as the optimizer with an initial learning rate of 0.001. The momentum was set to 0.9, and the weight decayed to 5e-4. The cosine annealing function was adopted as the learning rate adjustment strategy. The batch size was 48, and 4× Nvidia GeForce RTX 3090 was used for training. As for some of the hyperparameters, we set step size $\eta = 0.01$. And through the ablation study in Sec. 5, we set the parameter top $k = 20$, and the weight term $\alpha = 0.5$ in Eq. (7).

4.2. Robust Classification

In this section, we have compared the robustness of our method on benchmark datasets with several other SOTA methods, including FAN [60], RVT [36] and TAPADL [14], to demonstrate the superiority of our approach in terms of both computational efficiency and accuracy. The results are shown in Tab. 1. It is worth noting that we have primarily replicated the results of these methods, achieving consistency with their reports. Therefore, we directly adopt some of the results reported in their papers.

The results in Tab. 1 demonstrate that our method can be applied across various network architectures, consistently achieving significant improvements. Applying our method directly to the ViT-Base model improved the top-1 accuracy for clean images from 76.3% to 79.9%. Significant improvements were also observed in the robustness benchmark datasets IN-C, IN-P, IN-R, and IN-A. Replacing the backbone with a more advanced FAN [60] model yielded similar results. This demonstrates the versatility and gener-



(a) Image (b) TAPADL (c) Ours (d) Ground-Truth

Figure 5. Object detection results on RTTS and ExDark. From this visualization, we can observe that ours exhibits fewer missed detections.

alization capability of our method.

We also analyzed the parameters, GFLOPs, and inference time of each method. As shown in Tab. 1, it is worth noting that although the state-of-the-art FAN backbone network has fewer parameters and lower computational complexity, its non-standard modifications hinder parallel computation, significantly increasing inference time. Our experiments found that FAN requires 20.8ms for 100 inferences on a GPU, whereas the standard ViT-Base model only takes 4.8ms under the same conditions. Moreover, despite using the ViT-B backbone, our method outperforms the SOTA method TAPADL (FAN). For instance, the mCE (lower is better) of ImageNet-C improves from 43.7% to 41.4%. As a result, we conducted the subsequent experiments using the standard ViT-Base model. The detailed results under various degradation types in ImageNet-C can be found in the supplementary materials.

4.3. Object Detection in the Wild

We also evaluated our method on benchmark robustness object detection datasets, including COCO-C (images in MS-COCO [26] with corruptions similar to those in ImageNet-C), Foggy-CityScapes [43] (synthetic foggy images generated from Cityscapes), RTTS [24] (real-world foggy images), and ExDark [32] (real-world low-light images). We reported the overall mean Average Precision (mAP) values for object detection. We trained on the original clean images and tested on the degraded images for the synthetic benchmark datasets COCO-C and Foggy-CityScapes. As for the real-world dataset, since COCO shares categories with these datasets, we consistently trained on COCO and tested on RTTS and ExDark. During training, we used the fine-tuned model in Sec. 4.2 as the backbone and built the detection model (including neck and head) using `mmdet`[5]. We also generated some gradient perturbations images for training as described in Sec. 3 and introduced perturbations targeting the background class, as some regions of interest (ROI) are often misclassified as background.

Method	COCO-C	Foggy-Cityscapes	RTTS	ExDark
FAN [60]	39.0	37.5	43.1	52.6
TAPADL [14]	40.2	38.3	43.7	53.4
Ours	**41.7**	**39.1**	**44.5**	**54.3**

Table 2. Comparison with other methods on benchmark robustness object detection datasets. The best results are indicated in **bold**.

The results are listed in Tab. 2. We have also visualized some detection results on RTTS and ExDark in Fig. 5. From these experimental results, we observe that compared to others, our method exhibited fewer missed detections and achieved notable improvements in object detection as well.

5. Ablation Studies and Analysis

In Sec. 5.1, we first conducted several experiments to investigate the effectiveness of each component in our proposed method. Then, in Sec. 5.2, we studied the impact of hyperparameter k and α in Eq. (7), respectively. In Sec. 5.3, we discuss the reasons our method requires only a small number of samples and a few iterations to achieve surprisingly superior performance.

5.1. Effectiveness of Each Component

The two main components of our method are the gradient perturbation towards neighboring classes and the interclass distance-weighted loss. The exploration of perturbation types mainly includes the following three experiments.

Baseline: The baseline model is an original pre-trained vit-Base model without fine-tuning.

Standard Untargeted Adversarial Training: In this setting, perturbations are introduced without specifying a target class, aiming to increase the model's robustness by

Method	Perturbation	Weighted	IN (Acc) ↑	IN-C (mCE) ↓
Baseline	-	-	76.3	52.2
Adv.	untargeted	-	77.6 (+1.3)	48.7 (-3.5)
Nbr.	neighboring	-	79.1 (+2.8)	43.5 (-8.7)
Ours	neighboring	✓	**79.9** (+3.6)	**41.4** (-10.8)

Table 3. The ablation studies. "Adv." represents the standard untargeted adversarial training, and "Nbr." denotes the proposed neighboring class gradient perturbation without using weighted loss. The best results are indicated in **bold**.

Value	$k = 10$	$k = 20$	$k = 30$	$k = 40$	$k = 50$
Acc.	51.3	**51.9**	51.7	51.5	51.4

Table 4. The impact of different hyperparameter values k on unseen motion blur data. The best results are indicated in **bold**.

Value	$\alpha = 0.9$	$\alpha = 0.7$	$\alpha = 0.5$	$\alpha = 0.2$	$\alpha = 0.1$
Acc.	49.5	50.1	**51.9**	51.2	50.8

Table 5. The impact of different hyperparameter values α on unseen motion blur data. The best results are indicated in **bold**.

pushing predictions away from the true labels, as shown in Eq. (3). This is the conventional adversarial training where the perturbation forces the model to defend against arbitrary adversarial changes.

Neighboring Class Gradient Perturbation: In contrast, this method involves introducing perturbations through Eq. (5) that specifically direct the model's predictions toward semantically similar neighboring classes. These perturbations are more targeted, leveraging the fact that erroneous predictions due to distribution shifts often occur in closely related categories as described in Sec. 3.1.

We also investigated the effects of inter-class distance-weighted loss described in Sec. 3.3. Specifically, we compared the standard cross-entropy loss with the weighted loss defined in Eq. (8), respectively. Tab. 3 lists the results of all the ablation studies. It demonstrates that each module in our proposed method contributes to improving recognition performance and robustness.

5.2. Hyperparameter k and α

For the hyperparameter k and α, we assigned various values to these parameters and conducted a quick test on motion blur images at degradation severity level 5, ultimately selecting $k = 20$ and $\alpha = 0.5$. The results of the related experiments are shown in Tabs. 4 and 5.

5.3. Analysis of Faster Convergence

In our robust classification experiments, we found that fine-tuning the model for 10 epochs on all images from ImageNet-1k yielded similar results to using a subset of just 5000 images. This suggests that our method can enhance the robustness of pre-trained models using only a small number of samples and iterations. We suggest that the following two main factors contribute to this efficiency.

Reduced Optimization Space: Consider the standard optimization process where the model is trained to minimize the empirical risk,

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x}',y)\in D}[\mathcal{L}(\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}'),y]. \tag{9}$$

In standard adversarial training, as in Eq. 3, the model has to correct large errors (where $\boldsymbol{x}'$ is pushed far away from y), which can lead to slower convergence. In contrast, by constraining perturbations to the neighboring similar classes (where $\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{x}') = y_p$ is close to label y), we reduce the complexity of the optimization landscape. The constrained adversarial samples are "easier" to learn from, requiring fewer gradient updates. Mathematically, limiting adversarial perturbations to smaller regions makes the loss smoother and easier to optimize.

Refined Loss Penalty: The inter-class distance-weighted loss function further focuses the model's learning on high-risk areas. Given the weighting scheme, the loss focuses on categories with small inter-class distances, which are more important for enhancing model robustness. This prioritization forces the model to distinguish between classes that are easy to confuse efficiently. The model spends more capacity improving on these high-risk, high-weight regions, reducing the number of overall samples and iterations needed.

Together, these components enable the model to improve its robustness with fewer data and iterations, making the training process more efficient.

6. Conclusions

In this paper, we proposed a novel method that employs gradient-based perturbations targeting similar neighboring classes and an inter-class distance weighted loss to enhance model robustness in the wild. Comprehensive experiments demonstrate that our method strengthens the model's resilience to distributional shifts, as evidenced by improved performance on both synthetic and real-world degraded datasets. It consistently enhances robustness across diverse architectures, requiring minimal data and computational resources. Since perturbation samples from neighboring classes, which can be considered hard samples due to their susceptibility to misclassification, have been shown to enhance model robustness effectively, future work will explore more refined strategies like *hard sample mining* to guide the generation of such perturbation samples.

7. Acknowledgment

This work was supported in part by the Natural Science Foundation of China under Grant 62472336, Grant 61991451 and Grant 62302369.

References

[1] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? *Advances in neural information processing systems*, 34:26831–26843, 2021. 3

[2] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *International conference on learning representations*, 2018. 3

[3] Rich Caruana. Multitask learning. *Machine learning*, 28: 41–75, 1997. 3

[4] Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, and Yonghong Tian. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 458–467, 2021. 2

[5] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 7

[6] Sui Chenhong, Wang Ao, Zhou Shengwen, Zang Ankang, Pan Yunhao, Liu Hao, and Wang Haipeng. A survey on adversarial training for robust learning. *Journal of Image and Graphics*, 28(12):3629–3650, 2023. 3

[7] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019. 3

[8] Xiaohan Cui, Long Ma, Tengyu Ma, Jinyuan Liu, Xin Fan, and Risheng Liu. Trash to treasure: Low-light object detection via decomposition-and-aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1417–1425, 2024. 3

[9] Zhou Dawei, Xu Yibo, Wang Nannan, Liu Decheng, Peng Chunlei, and Gao Xinbo. Generalized adversarial defense against unseen attacks: a survey. *Journal of Image and Graphics*, 29(07):1787–1813, 2024. 3

[10] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[11] Zhipeng Du, Miaojing Shi, and Jiankang Deng. Boosting object detection with zero-shot day-night domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12666–12676, 2024. 3

[12] Haoyang Fang, Boran Han, Shuai Zhang, Su Zhou, Cuixiong Hu, and Wen-Ming Ye. Data augmentation for object detection via controllable diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1257–1266, 2024. 3

[13] Yong Guo, David Stutz, and Bernt Schiele. Improving robustness of vision transformers by reducing sensitivity to patch corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4108–4118, 2023. 6

[14] Yong Guo, David Stutz, and Bernt Schiele. Robustifying token attention for vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17557–17568, 2023. 2, 3, 6, 7

[15] Himanshu Gupta, Oleksandr Kotlyar, Henrik Andreasson, and Achim J Lilienthal. Robust object detection in challenging weather conditions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7523–7532, 2024. 3

[16] Khurram Azeem Hashmi, Goutham Kallempudi, Didier Stricker, and Muhammad Zeshan Afzal. Featenhancer: Enhancing hierarchical features for object detection and beyond under low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6725–6735, 2023. 3

[17] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 4, 6

[18] 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. 2

[19] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 3, 6

[20] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 6

[21] Zeyi Huang, Andy Zhou, Zijian Ling, Mu Cai, Haohan Wang, and Yong Jae Lee. A sentence speaks a thousand images: Domain generalization through distilling clip with language guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11685–11695, 2023. 3

[22] Khawar Islam, Muhammad Zaigham Zaheer, Arif Mahmood, and Karthik Nandakumar. Diffusemix: Label-preserving data augmentation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27621–27630, 2024. 3

[23] Insoo Kim, Seungju Han, Ji-won Baek, Seong-Jin Park, Jae-Joon Han, and Jinwoo Shin. Quality-agnostic image recognition via invertible decoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12257–12266, 2021. 3

[24] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2018. 7

[25] Jinlong Li, Runsheng Xu, Jin Ma, Qin Zou, Jiaqi Ma, and Hongkai Yu. Domain adaptive object detection for autonomous driving under foggy weather. In *Proceedings of*

the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 612–622, 2023. 3

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 7

[27] Ding Liu, Bihan Wen, Xianming Liu, Zhangyang Wang, and Thomas S Huang. When image denoising meets high-level vision tasks: A deep learning approach. *arXiv preprint arXiv:1706.04284*, 2017. 3

[28] Hongmin Liu, Fan Jin, Hui Zeng, Huayan Pu, and Bin Fan. Image enhancement guided object detection in visually degraded scenes. *IEEE transactions on neural networks and learning systems*, 2023. 3

[29] Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. Image-adaptive yolo for object detection in adverse weather conditions. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1792–1800, 2022. 3

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[31] Zicheng Liu, Siyuan Li, Di Wu, Zihan Liu, Zhiyuan Chen, Lirong Wu, and Stan Z Li. Automix: Unveiling the power of mixup for stronger classifiers. In *European Conference on Computer Vision*, pages 441–458. Springer, 2022. 2

[32] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019. 7

[33] Raphael Gontijo Lopes, Dong Yin, Ben Poole, Justin Gilmer, and Ekin D Cubuk. Improving robustness without sacrificing accuracy with patch gaussian augmentation. *arXiv preprint arXiv:1906.02611*, 2019. 2

[34] Andong Lu, Tianrui Zha, Chenglong Li, Jin Tang, Xiaofeng Wang, and Bin Luo. Nighttime person re-identification via collaborative enhancement network with multi-domain learning. *arXiv preprint arXiv:2312.16246*, 2023. 3

[35] Aleksander Madry. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2, 3

[36] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 12042–12051, 2022. 3, 6

[37] 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. *arXiv preprint arXiv:1907.07484*, 2019. 3

[38] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[39] Wan Peng, Hu Cong, and Wu Xiaojun. Multi-domain feature mixup boosting adversarial examples transferability method. *Journal of Image and Graphics*, 29(12):3670–3683, 2024. 3

[40] Huafeng Qin, Xin Jin, Hongyu Zhu, Hongchao Liao, Mounîm A El-Yacoubi, and Xinbo Gao. Sumix: Mixup with semantic and uncertain information. *arXiv preprint arXiv:2407.07805*, 2024. 2

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[42] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. 6

[43] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 7

[44] Aaditya Singh, Kartik Sarangmath, Prithvijit Chattopadhyay, and Judy Hoffman. Benchmarking low-shot robustness to natural distribution shifts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16232–16242, 2023. 3

[45] Taeyoung Son, Juwon Kang, Namyup Kim, Sunghyun Cho, and Suha Kwak. Urie: Universal image enhancement for visual recognition in the wild. In *European Conference on Computer Vision*, pages 749–765. Springer, 2020. 3

[46] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In *Proceedings of the European conference on computer vision (ECCV)*, pages 631–648, 2018. 3

[47] Weimin Tan, Bo Yan, and Bahetiyaer Bare. Feature super-resolution: Make machine see more clearly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3994–4002, 2018. 3

[48] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 6

[49] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999. 1

[50] 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*, pages 3219–3229, 2023. 3

[51] Haotao Wang, Chaowei Xiao, Jean Kossaifi, Zhiding Yu, Anima Anandkumar, and Zhangyang Wang. Augmax: Adversarial composition of random augmentations for robust training. *Advances in neural information processing systems*, 34: 237–250, 2021. 2

[52] Yang Wang, Yang Cao, Zheng-Jun Zha, Jing Zhang, and Zhiwei Xiong. Deep degradation prior for low-quality image

classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11049–11058, 2020. 3

[53] Yang Wang, Tieyong Cao, Jibin Yang, Yunfei Zheng, Zheng Fang, and Xiaotong Deng. A perturbation constraint related weak perceptual adversarial example generation method. *Journal of Image and Graphics*, 27(7):2287–2299, 2022. 3

[54] Zhou Yang, Weisheng Dong, Xin Li, Mengluan Huang, Yulin Sun, and Guangming Shi. Vector quantization with self-attention for quality-independent representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24438–24448, 2023. 3

[55] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 2

[56] Hongyi Zhang. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2

[57] Yuwei Zhang, Yan Wu, Yanming Liu, and Xinyue Peng. Cpa-enhancer: Chain-of-thought prompted adaptive enhancer for object detection under unknown degradations. *arXiv preprint arXiv:2403.11220*, 2024. 3

[58] Yin Zhang, Yongqiang Zhang, Zian Zhang, Man Zhang, Rui Tian, and Mingli Ding. Isp-teacher: Image signal process with disentanglement regularization for unsupervised domain adaptive dark object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7387–7395, 2024. 3

[59] Peng Zhenbang, Zhang Yu, Dang Yi, et al. Review of physical adversarial attacks against visual deep learning models. *Journal of Image and Graphics*, 2024. [Online]. 3

[60] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pages 27378–27394. PMLR, 2022. 2, 3, 6, 7