VITA: A Multi-Source Vicinal Transfer Augmentation Method for Out-of-Distribution Generalization

Minghui Chen[1], Cheng Wen[2], Feng Zheng[1*], Fengxiang He[3], Ling Shao[4]

[1] Department of Computer Science and Engineering, Southern University of Science and Technology
[2] The University of Sydney [3] JD Explore Academy
[4] National Center for Artificial Intelligence, Saudi Data and Artificial Intelligence Authority, Riyadh, Saudi Arabia
ming_hui.chen@outlook.com, cwen6671@uni.sydney.edu.au, f.zheng@ieee.org, hefengxiang@jd.com, ling.shao@ieee.org

Abstract

Invariance to diverse types of image corruption, such as noise, blurring, or colour shifts, is essential to establish robust models in computer vision. Data augmentation has been proven to improve robustness against common corruptions. However, existing augmentation strategies for boosting corruption robustness produce samples that deviate significantly from the underlying data manifold. As a result, performance is skewed toward certain types of corruption. To address this issue, we propose a multi-source vicinal transfer augmentation (VITA) method for generating diverse on-manifold samples. The proposed VITA consists of two complementary parts: tangent transfer and integration of multi-source vicinal samples. The tangent transfer creates initial augmented samples for improving corruption robustness. The integration employs a generative model to characterize the underlying manifold built by vicinal samples, facilitating in the generation of on-manifold samples. Our proposed VITA outperforms other state-of-the-art augmentation methods significantly, as demonstrated by extensive experiments on corruption benchmarks.

1 Introduction

Existing computer vision systems are not as robust as the human vision system (Recht et al. 2018; Hendrycks and Dietterich 2019). The human vision system is not confused by a wide range of naturally occurring corruptions such as noise, blurring, and pixelation, as well as unexpected combinations of these. However, current deep models (Krizhevsky, Sutskever, and Hinton 2012; Xie et al. 2017) trained on clean images typically perform substantially worse when confronted with corrupted images (Geirhos et al. 2018; Hendrycks and Dietterich 2019). Achieving this level of robustness is a primary objective of a variety of computer vision tasks.

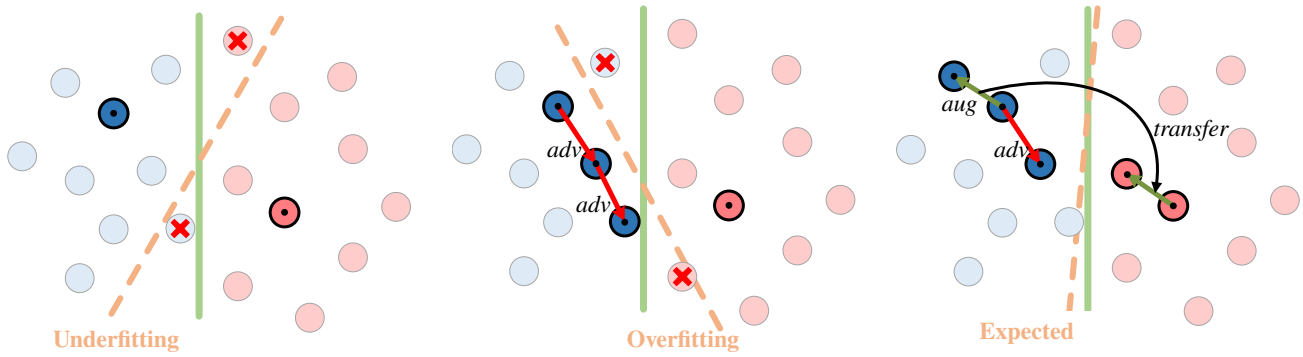
The most effective and commonly used method for improving corruption robustness is data augmentation. Data augmentation is a technique in which training samples undergo label-preserving transformations (Dao et al. 2019), formalized by the vicinal risk minimization (VRM) principle (Chapelle et al. 2000). Employing adversarial examples (Goodfellow, Shlens, and Szegedy 2015; Madry et al. 2018) or transformation strategies (Devries and Taylor 2017; Zhang et al. 2018; Yun et al. 2019) based on human priors brings limited benefit to corruption robustness. This is due to the fact that these two methods of augmentation fail to provide sufficiently diverse vicinal samples. While advanced methods based on generative models (Geirhos et al. 2019; Rusak et al. 2020; Hendrycks et al. 2020a) and combination strategies (Cubuk et al. 2019; Hendrycks et al. 2020b) are capable of producing diverse samples, they frequently generate samples that deviate severely from the data manifold. These samples impair the classifier's ability to estimate the underlying data manifold accurately, resulting in performance degradation and bias towards specific corruptions.

To mitigate this issue, we propose a multi-source vicinal transfer augmentation (VITA) method for generating diverse on-manifold samples. The proposed VITA is composed of two components: the tangent transfer and the integration of multi-source vicinal samples. First, we leverage vicinal differences to approximate the manifold tangents to acquire initial augmented samples. Subsequent experiments show that these weakly augmented samples effectively improve corruption robustness. Second, we use a generative model to characterize the underlying data manifold constructed by weakly augmented samples (*e.g.*, samples rotated by 5 degrees) and adversarial examples. This is to ensure that a diverse set of samples is generated while avoiding significant deviance from the data manifold (Bengio et al. 2013). Detailed experiments confirm that our VITA can significantly improve corruption robustness and encourage a balanced performance on corrupted datasets.

In summary, our key contributions are as follows:

- To address the uneven performance toward various corrupted images, we propose a multi-source vicinal transfer augmentation (VITA) method for generating diverse on-manifold samples.
- We introduce tangent transfer that enforces the local invariance of the classifier, which facilitates the discovery of shared structures in the tangent planes.
- We design an integration module of multi-source vicinal samples that constructs a proper data manifold and is shown to effectively generate on-manifold samples.
- Our proposed VITA achieves state-of-the-art performances on corruption benchmark datasets CIFAR-10-C,

*Corresponding Author.



(a) Lack of vicinity sampling (b) Off-manifold samples hurts (c) Vicinal difference transfer

Figure 1: Diverse on-manifold samples boost performance. **(a)** shows an underfitting phenomenon caused by the lack of sufficiently diverse vicinal samples. **(b)** takes a strong adversarial attack as an example to show the performance degradation caused by off-manifold samples. Classifiers tend to overfit these off-manifold samples and fail to construct a proper underlying manifold. **(c)** demonstrates the benefit of tangent transfer. While maintaining sample variety, it will not generate severe out-of-manifold samples. For a better understanding of samples that depart from the manifold, see our appendix for a three-dimensional schematic diagram.

CIFAR-100-C, and ImageNet-C. Meanwhile, we demonstrate VITA also boosts adversarial robustness.

2 Related Work

Corruption Robustness. The human visual system is not easily defrauded by data with various forms of corruption, such as snowflakes, blurring, pixelation, or their combinations. In contrast, most current deep learning models suffer from severe performance degradation under corrupted data (Vasiljevic, Chakrabarti, and Shakhnarovich 2016; Dodge and Karam 2017; Azulay and Weiss 2019). For example, (Geirhos et al. 2018) reveal that deep neural networks trained on one type of corruption (*e.g.*, salt-and-pepper noise) cannot recognize another unseen type of corruption (*e.g.*, uniform white noise), even though these two kinds of corruptions are indistinguishable to humans. Currently, research on improving corruption robustness mainly focuses on domain adaptation (*e.g.*, additional operations on the normalization layer) (Schneider et al. 2020; Tang et al. 2021), adversarial perturbations (Hendrycks and Dietterich 2019) and data augmentation (Zhang et al. 2018; Hendrycks et al. 2020b,a; Kamann and Rother 2020; Rusak et al. 2020). To evaluate the corruption robustness of models, (Hendrycks and Dietterich 2019) introduced three comprehensive benchmarks, CIFAR-10-C, CIFAR-100-C and ImageNet-C, for unseen corruption robustness. Since then, similar datasets on common corruptions have also been proposed in the field of object detection (PASCAL-C, COCO-C and Cityscapes-C) (Michaelis et al. 2019) and semantic segmentation (Kamann and Rother 2020). These benchmarks demonstrate that the generalization ability of many advanced models under corrupted input still needs to be further improved.

Data Augmentation. Data augmentation is one of the most widely studied and effective techniques to improve the corruption robustness of models. For example, Mixup (Zhang et al. 2018) is a simple augmentation strategy that performs a linear interpolation between two different classes of samples. Although it was not specifically proposed to improve corruption robustness, its performance is significantly better than other commonly used data augmentation methods (Dodge and Karam 2017; Yun et al. 2019; Cubuk et al. 2019). Another effective data preprocessing method called AugMix (Hendrycks et al. 2020b) obtains advanced performance on CIFAR-10-C, CIFAR-100-C and ImageNet-C. It utilizes a formulation to mix multiple augmented images and adopts a Jensen-Shannon Divergence consistency loss. Further, (Rusak et al. 2020) demonstrate that data augmented with Gaussian noise can serve as a simple yet very strong baseline for defending against common corruptions.

3 Proposed Method

In this section, we present the proposed method VITA in detail. Our proposed VITA involves two stages. The first stage, tangent transfer in 3.1 yields initial augmented samples for improving corruption robustness. The second stage, multi-source sample integration in 3.2 generates diverse and on-manifold samples to further improve corruption robustness. In 3.3, we describe how to train with VITA samples.

3.1 Tangent Transfer

Exploiting Shared Manifold Structure. (Bengio and Monperrus 2004) used the shared structure of the tangent space of the manifold to mitigate the curse of dimensionality in the previous local manifold learning algorithms. Generally, in many real-world contexts, there is not just one global manifold but a large number of manifolds that share something about their structure (Bengio and Monperrus 2004; Lasserre, Bishop, and Minka 2006). However, existing research on data augmentation ignores leveraging this characteristic of the data manifold. A simple example is transformations in the image (rotation, lighting, blurring *etc.*). There is one manifold for each transformation type. If there are only a few samples of a specific type of transformation during training, it is hard for models to learn a proper

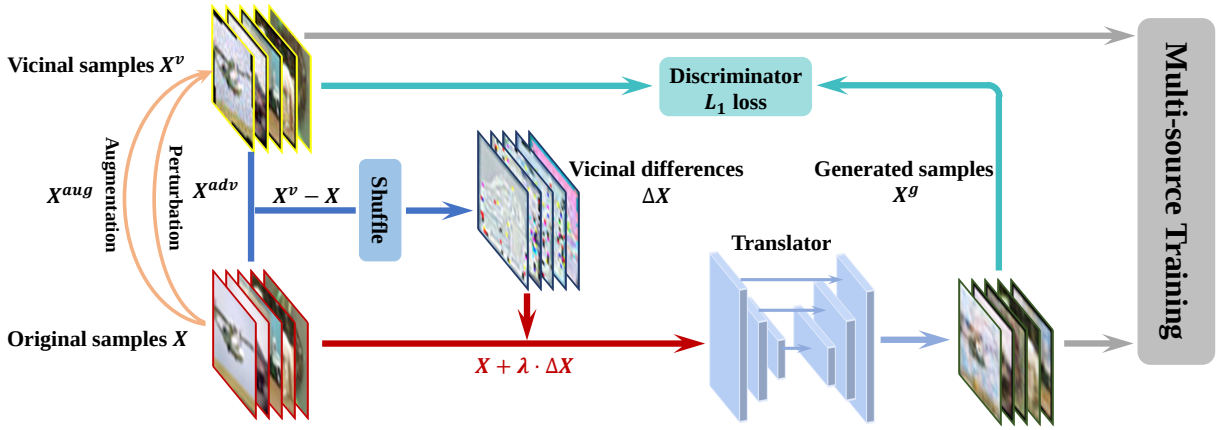


Figure 2: Our proposed VITA includes tangent transfer and multi-source integration. We adopt vicinal difference (ΔX) as an approximation of the manifold tangent and use $X + \lambda \cdot \Delta X$ as an initial augmented sample. Our generative model includes a $pix2pix$ model, which is designed to generate diverse on-manifold samples. The goal of multi-source integration is to learn, based on dataset $\mathcal{D} = \{X^v, Y\}$ ($X^v = \{X^{\text{aug}}, X^{adv}\}$), an embedding that imitates the generation process of vicinal samples $P(x^v|x, \delta x)$, where x is an original samples and $\delta x \in \Delta X (\Delta X = X^v - X)$. Note that δx is a transferred vicinal difference. In our robust multi-source training, models are trained with multi-source samples and samples from a well-trained translator.

data manifold. As shown in Fig. 1, this leads models to directly memorize these special cases instead of generalizing (*i.e.*, yields the high complexity or unsmooth data manifold). This, we believe, is the core reason for the model's poor and uneven performance against various forms of corruption. If the learned structure of the data manifold is shared, models can more accurately characterize a smoother data manifold.

In this work, we use vicinal differences as a rough approximation of manifold tangents. Transferring vicinal differences enforces the local invariance of the classifier and encourages the classifier to discover shared structures in the tangent planes at different positions. We argue that approximating the tangent direction of the manifold with the vicinal difference of high-dimensional space can yield more diverse samples, which differs from a model-sensitive measure (Simard et al. 1991, 1996; Lopes et al. 2020). Subsequent experiments demonstrated the effectiveness of introducing shuffled vicinal differences.

In specific, vicinal differences are obtained by subtracting the original samples from the vicinal samples. Here, vicinal samples are crafted through diverse data augmentation operations and adversarial attack methods. The tangent transfer is realized by adding shuffled vicinal differences.

Transformation Guided by Priors. First, we introduce several hand-crafted methods for weakly augmented samples used in our work, including rotation, shearing, translating, cropping and scaling (detailed settings of these augmentations see appendix). To make the augmented sample more consistent at input level, we need to control the intensity of the change: $||\delta x||_2 = ||x^{\text{aug}} - x||_2 < \epsilon_2$, where ϵ is a hyper-parameter ($\epsilon_2 = 0.5$ by default), x is an original sample and x^{aug} is an augmented sample.

Harvesting Adversarial Perturbations. In contrast to data augmentation, crafting adversarial examples makes use of existing trained models instead of priors from human experience. Thus, we employ several algorithms to generate adversarial examples, including the fast gradient sign method (FGSM) (Goodfellow, Shlens, and Szegedy 2015), projected gradient descent (PGD) (Madry et al. 2018), momentum iterative method (Dong et al. 2018), C&W method (Carlini and Wagner 2017) and Elastic-Net Attack (Chen et al. 2018) (detailed hyper-parameter settings see appendix). For different well-trained models, the adversarial examples of a given seed x will vary because of the diverse loss functions and hypothetical spaces. However, they all reflect the local structure of the seed x from the different perspectives. For adversarial examples, we also control the magnitude of perturbations. For all datasets, we use the $\ell_\infty-$ and ℓ_2- adversarial attack methods with fixed budget of $\epsilon_\infty = 0.031$ and $\epsilon_2 = 0.5$.

3.2 Multi-Source Samples Integration

Utilizing Differences via a Generative Model In the following experiments, we integrate all the crafted samples, which include the multi-source vicinal examples from data augmentation and adversarial perturbations, into a dataset $\mathcal{D} = \{X^{\text{aug}}, X^{\text{adv}}, Y\}$ and treat them equivalently. Note that $x^v \in \{X^{\text{aug}}, X^{adv}\}$ is the input data and $y \in Y$ is the label of x^v. Our goal is to learn, based on dataset $\mathcal{D}$, an embedding that imitates the generation process of vicinal samples $P(x^v|x, \delta x)$ given x and δx. Note that $\delta x \in \Delta X$ is a transferred vicinal difference. To this end, we use an adversarial loss (Goodfellow et al. 2014) to implement the embedding and build two basic models: a sample-to-sample translation model T and a discriminative model D.

Mapping from input distribution to output distribution in

high-dimensional space is challenging (Zhu et al. 2017). Thus, we start with the pix2pix framework (Isola et al. 2017), which has previously been shown to produce high-quality results for various image-to-image translation tasks. In our appendix, we also elaborate on the ablation experiments related to the translator, discriminator and the impact of using a more complex framework such as BicycleGAN (Zhu et al. 2017).

Translator. The sample with vicinal difference may not fall on the local data manifold because of the complex high-dimensional data space. Thus, we need to build a translator to embed the raw intermediate product $x + \delta x$ into the on-manifold vicinal sample $x^g = T(x + \delta x)$. As shown in Fig. 2, inputs to our translator are the samples added with shuffled vicinal differences, and outputs are desired samples on the data manifold. By default, when training the translator, we include the same proportion of augmented and adversarial data. Our translator takes advantage of a U-Net structure (Ronneberger, Fischer, and Brox 2015), which enables the transmission of hierarchical information across a network by skipping layer connections. In our experiments, the U-Net structure has shown its effectiveness to preserve the vicinal differences. For more details on the superiority of the U-Net architecture, see our ablation experiments in the appendix.

Discriminator. We denote our discriminator as D, and $D(x^g)$ indicates the probability that a generated sample x^g comes from the real vicinity. In our network, we use a 1×1 PatchGAN (Isola et al. 2017) discriminator by default. PatchGAN discriminator is a type of discriminator for generative adversarial networks which only penalizes structures at the scale of the local image patches. The PatchGAN discriminator tries to classify whether each patch in an image is real or fake.

Objective Function. The embedding of the translator and discriminator is configured using neural nets. We describe the main part of loss below.

$$\mathcal{L}_{\text{GAN}}(T, D) = \mathbb{E}_{x,x^v \sim p(x,x^v)}[\log(D(x, x^v))] + \mathbb{E}_{x \sim p(x), \delta x \sim p(\delta x)}[\log(1 - D(x, T(x + \delta x)))] \quad (1)$$

Here, $p(x, x^v)$ represents the joint distribution of the original samples distribution x and the vicinal samples x^v distribution, $p(x)$ denotes the distribution of x, and $p(\delta x)$ represents the the distribution of vicinal differences.

3.3 Robust Training Process

We argue that multi-source samples can provide more vicinal information, as shown in Fig. 1. In order to make full use of multi-source samples, we divide samples into three categories according to their source, namely the weakly augmented samples, the samples added with shuffled adversarial perturbations, the samples generated via VITA. During training, we split the samples in each batch into weakly augmented samples (25%), shuffled perturbations samples (25%) and generated samples via VITA (50%). Among the generated samples, half of the vicinal differences for the translator come from weakly augmented samples, and half of them are generated from shuffled adversarial perturbations. More specifically, we apply a Jensen-Shannon Divergence consistency loss as a regularization term to enforce a consistent embedding by the classifier across further diverse augmentation. More details on our robust training process can be found in our supplementary materials.

4 Experiments

Dataset. CIFAR-10 (10 categories) and CIFAR-100 (100 categories) both contain small $32 \times 32 \times 3$ colour images, with 50k for training and 10k for testing. The ImageNet (Deng et al. 2009) dataset includes 1,000 classes and contains approximately 1.2 million images annotated according to the WordNet hierarchy. To evaluate the corruption robustness of models, we conduct experiments on the CIFAR-10-C, CIFAR-100-C and ImageNet-C datasets (Hendrycks and Dietterich 2019). These datasets are constructed by corrupting the original images from the CIFAR-10, CIFAR-100 and ImageNet test sets. Specifically, the CIFAR-10-C, CIFAR-100-C, and ImageNet-C datasets consist of 15 types of algorithmically generated corruptions from noise, blur, weather, and digital categories. Each type of corruption has five levels of severity, resulting in 75 distinct corruptions. Since these datasets are used to measure model performance under data shifts, we take care not to introduce the 15 corruptions into the VITA process and robust training procedure.

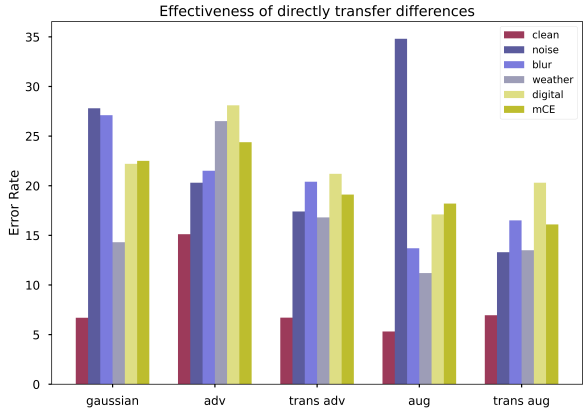


Figure 3: Effectiveness of directly transferring differences. We compare the corruption robustness of models trained with different augmentation methods, including Gaussian noise (Gaussian), adversarial perturbations (adv), transferred perturbations (trans adv), weak augmentation (aug), transferred differences from weak augmentation (trans aug). We evaluate corruption robustness (error rate, the lower, the better) on CIFAR-10-C with an AllConvNet architecture. Here, *clean* indicates the performance on clean (uncorrupted) images, *noise/blur/weather/digital* is robustness towards the corresponding corruption types and *mCE* is mean corruption error for all types of corruption. As seen, transferring vicinal differences improves robustness significantly.

Table 1: Mean corruption error (mCE) on CIFAR-10-C and CIFAR-100-C. Average classification error as percentages. Across several architectures, our method obtains obvious enhancement in corruption robustness. Specifically, we reduced the error rate of corrupted data to 8.9 on ResNeXt.

		Standard	Cutout	Mixup	CutMix	AutoAug	Adv Train	AugMix	ME-ADA	VITA
CIFAR-10-C	AllConvNet	30.8	32.9	24.6	31.3	29.2	28.1	15.0	21.8	**10.6**
	DenseNet	30.7	32.1	24.6	33.5	26.6	27.6	12.7	23.1	**9.7**
	WideResNet	26.9	26.8	22.3	27.1	23.9	26.2	11.2	16.7	**9.5**
	ResNeXt	27.5	28.9	22.6	29.5	24.2	27.0	10.9	16.6	**8.9**
Mean		29.0	30.2	23.5	30.3	26.0	27.2	12.5	19.5	**9.7**
CIFAR-100-C	AllConvNet	56.4	56.8	53.4	56.0	55.1	56.0	42.7	48.8	**36.3**
	DenseNet	59.3	59.6	55.4	59.2	53.9	55.2	39.6	52.2	**35.4**
	WideResNet	53.3	53.5	50.4	52.9	49.6	55.1	35.9	47.2	**34.4**
	ResNeXt	53.4	54.6	51.4	54.1	51.3	54.4	34.9	42.7	**31.5**
Mean		55.6	56.1	52.6	55.5	52.5	55.2	38.3	47.7	**34.4**

Metric. The clean error is the usual classification error on uncorrupted test images. In terms of measuring corruption robustness, we use mean error at five different intensities or levels of severity, i.e. $1 \leq s \leq 5$. Let $E_{c,s}$ denote the test error of corrupted images from corruption type c and under severity level s. For CIFAR datasets, we use the mean corruption error (mCE) over fifteen corruptions and five severities, i.e. $mCE = 1/75 \sum_{c=1}^{15} \sum_{s=1}^{5} E_{c,s}$. For ImageNet, we follow the convention of normalizing the corruption error by the corruption error of AlexNet (Krizhevsky, Sutskever, and Hinton 2012), i.e. $CE_c = \sum_{s=1}^{5} E_{c,s} / \sum_{s=1}^{5} E_{c,s}^{AlexNet}$. The mean of the 15 corruption errors gives us the $mCE = 1/15 \sum_{c=1}^{15} CE_c$.

4.1 Effectiveness of Transferring Differences

Setup. Our verification experiment is based on the All-Conv (Springenberg et al. 2015) network, trained on the clean CIFAR-10 dataset and evaluated on the CIFAR-10-C dataset. We mainly compare the impact of five different inputs on the corruption robustness. These inputs including samples added with Gaussian noise (*Gaussian*, mean value 0, standard deviation 0.5), adversarial perturbations (*adv*), transferred perturbations (*trans adv*). Training with weakly augmented samples (*aug*), and samples with transferred differences from weakly augmented samples (*trans aug*) are also included. In our supplementary materials, we conduct an ablation study on the types of augmentation and perturbations. And it further reveals the effectiveness of transferring difference on alleviating the model's sensitivity to certain types of local direction.

Results. Our first experiment verifies the validity of transferring vicinal differences. Our main findings are as follows. The first and most important thing is that transferring vicinal differences can improve the corruption robustness of, as shown in Fig. 3. In particular, transferring vicinal differences from weakly augmented samples can greatly improve robustness toward common corruptions. Also, the improvement brought by transferring vicinal differences is significantly greater than the addition of Gaussian noise. Interestingly, we discover that transferring adversarial perturbations improves robustness to noise corruption more effectively than adding Gaussian noise.

4.2 Effectiveness of VITA

In this part, we verify the effectiveness of our vicinal information fusion framework, that is, adding samples generated by VITA to the multi-source robust training process (as described in 3.3).

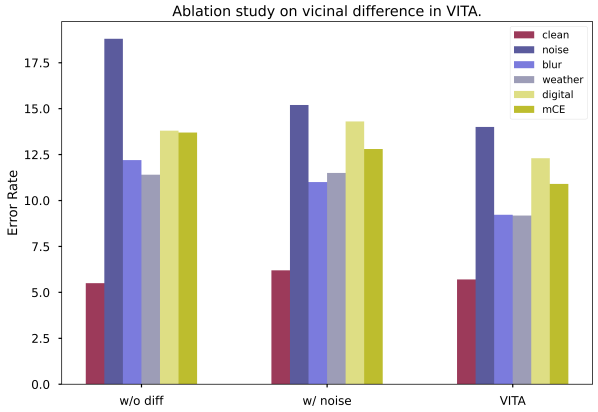


Figure 4: Ablation study on vicinal differences in VITA, demonstrating the necessity of transferred vicinal differences as input to a translator. We evaluate corruption robustness (error rate, the lower, the better) on CIFAR-10-C with an AllConvNet architecture. Here, *w/o diff* is the translator trained and inferred without transferred differences, and *w/ noise* is the translator trained and inferred with the addition of Gaussian noise. Training process is the same setting as for default robust training process (50% gen + 25 % adv + 25% aug). As seen, a translator trained with Gaussian noise or without vicinal differences (*i.e.* merely original data) performs worse than VITA.

CIFAR Training Settings. In the following experiments, we choose the same network architectures as AugMix

Table 2: Clean error, mean corruption error (mCE) and all types of corruption error rate values for various methods on ImageNet-C. We compare against other data augmentation methods for improving the corruption robustness. Our proposed VITA hugely improves corruption robustness and achieves balanced performance toward different corrupted images.

		Noise			Blur				Weather				Digital				
Network	Clean	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG	**mCE**
Standard	23.9	79	80	82	82	90	84	80	86	81	75	65	79	91	77	80	80.6
Patch Uniform	24.5	67	68	70	74	83	81	77	80	74	75	62	77	84	71	71	74.3
AutoAug	22.8	69	68	72	77	83	80	81	79	75	64	56	70	88	57	71	72.7
MaxBlur pool	23.0	73	74	76	74	86	78	77	77	72	63	56	68	86	71	71	73.4
SIN	27.2	69	70	70	77	84	76	82	74	75	69	65	69	80	64	77	73.3
AugMix	**22.4**	65	66	67	70	80	66	66	75	72	67	58	58	79	69	69	68.4
DeepAug	23.3	49	50	47	59	73	65	76	64	60	58	51	61	76	48	67	60.4
ANT	23.9	**39**	**40**	**39**	68	78	73	77	71	66	68	55	69	79	63	64	63.3
VITA	25.4	40	41	41	**47**	**61**	**51**	**59**	**57**	**58**	**55**	**49**	**47**	**69**	**44**	**62**	**52.1**

(Hendrycks et al. 2020b), including All Convolutional Network (Springenberg et al. 2015), DenseNet-BC ($k = 2, d = 100$) (Huang et al. 2017), 40-2 WideResNet (Zagoruyko and Komodakis 2016) and ResNeXt-29 (32×4) (Xie et al. 2017). We use stochastic gradient descent with an initial learning rate of 0.1 and *ReduceOnPlateau* scheduler. We train all architectures over 150 epochs.

CIFAR Results. We perform a comprehensive evaluation to compare with a total of 7 advanced augmentation methods, including Cutout (Devries and Taylor 2017), Mixup (Zhang et al. 2018), CutMix (Yun et al. 2019), AutoAug (Cubuk et al. 2019), adversarial training (Adv Train) (Carlini and Wagner 2017), AugMix (Hendrycks et al. 2020b), ME-ADA (Zhao et al. 2020). Compared to the standard data augmentation baseline (mean of four different architectures), our approach achieves 19.3% lower mCE as shown in Fig. 1. Compared to AugMix, which is the current state of the art on CIFAR-10-C and CIFAR-100-C, our method obtains significant performance improvement under various network architectures. Specifically, we achieve a 4.4% (CIFAR-10-C) and 6.4% (CIFAR-100-C) performance improvement in mCE under the AllConvNet compared with AugMix.

ImageNet Training Settings. We use ResNet-50 as the backbone of our model trained on ImageNet. The training scheme follows AugMix; that is, we apply a small learning rate for the first five epochs to warm up the training and then apply a decayed learning rate for the remaining epochs. In addition to AugMix with standard training, we also compare our method with DeepAug (Hendrycks et al. 2020a), stylized image training (SIN) (Geirhos et al. 2019) and adversarial noise training (ANT) (Michaelis et al. 2019). Stylized image training refers to the method in which the model is not only trained on the original dataset but also on stylized ImageNet samples (Geirhos et al. 2019).

ImageNet Results. We perform a large-scale evaluation to compare with a total of 7 advanced augmentation methods, including Patch Uniform (Lopes et al. 2019), AutoAug (Cubuk et al. 2019), MaxBlur pool (Zhang 2019), SIN (Geirhos et al. 2019), AugMix (Hendrycks et al. 2020b), DeepAug (Hendrycks et al. 2020a), ANT (Rusak et al. 2020). Although our method has a slight drop in accuracy on clean samples, it achieves 52.1% mCE as shown in Fig. 2, compared to AugMix with 68.4% and ANT with 63.3%. Moreover, even without Gaussian noise in the training phase, our method's performance in terms of noise corruption is close to ANT's (adversarial noise training). This shows the ability of VITA to promote balanced performance on different corruption types.

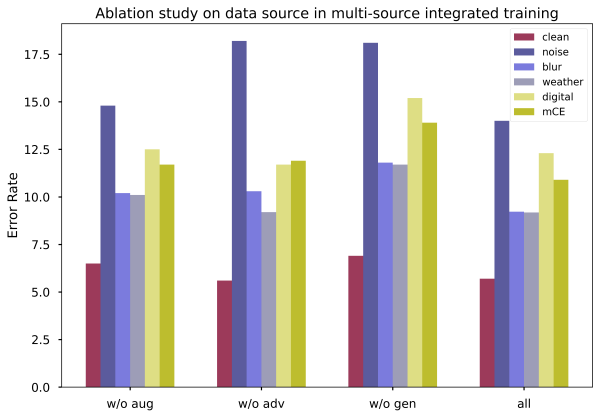


Figure 5: Ablation study on the data source in multi-source robust training reveals the importance of training with samples from VITA. We evaluate corruption robustness (error rate, the lower the better) on CIFAR-10-C with an AllConvNet architecture. Here, *w/o aug/adv/gen* indicate that specific types of samples lack in our multi-source robust training. For example, we train with 50% of samples with shuffled perturbations and 50% samples generated by VITA during training under the *w/o aug* setting. Obviously, when robust training does not include samples generated by VITA, corruption robustness decreases significantly.

4.3 Ablation Study on Vicinal Differences in VITA

Setup. In this part, we investigate the effects of inputs to the translator on the quality of the generated sample (i.e. the impact on improving the robustness of corruption). We con-

Table 3: Evaluations (test accuracy) of deep models (WRN-34-10) on the CIFAR-10 dataset. Results of TRADES ($\beta = 1.0$ and 6.0) are reported in (Zhang et al. 2019). Results of FAT for TRADES are reported in (Zhang et al. 2020). Our proposed framework of VITA and multi-source integrated training can also improve the adversarial robustness of the model.

Defense	Natural	FGSM	PGD-20	C&W$_\infty$	PGD-100	AutoAttack
TRADES ($\beta = 1.0$)	88.64	56.38	49.14	-	-	-
FAT for TRADES	**89.94**	61.00	49.70	49.35	48.35	47.22
VITA for Adv. Training	89.35	**68.02**	**52.33**	**50.21**	**50.04**	**48.38**
TRADES ($\beta = 6.0$)	84.92	61.06	56.61	54.47	55.47	53.08
FAT for TRADES	**86.60**	61.97	55.98	54.29	55.34	53.27
VITA for Adv. Training	85.75	**67.99**	**57.63**	**55.32**	**56.87**	**54.35**

duct our ablation study on CIFAR-10-C with an AllConvNet architecture. We add two sets of control experiments: one with the translator trained and inferred without vicinal differences (*w/o diff*) and one with Gaussian noise (*w/ noise*).

Results. As illustrated in Fig. 4, robust training with samples generated by the translator trained without vicinal differences is less robust against corruption, particularly noise corruption. Although training with samples from a noise-added translator improves performance against noise corruption, it remains significantly poorer than VITA on other types of corruption.

4.4 Ablation Study on Data Source in Training

Setup. In our multi-source robust training, we have three types of samples ($X^{aug}, X^{adv}, X^{gen}$). This part aims to determine which data source or combination of data sources most significantly contributes to the enhancement of corruption robustness and performance on a clean test set (for more ablation experiments on data sources and dataset size, see appendix). In Fig. 5, *w/o aug* (*w/o adv/gen* are similar) indicates that we only remove augmented samples during training (i.e. 50% *adv* + 50% *gen*). It should be emphasized that our data source ablation experiment is conducted during the multi-source robust training stage, and we still use multi-source samples when training the translator.

Results. Firstly, as can be seen from Fig. 5, the samples generated by VITA play an important role in improving the corruption robustness of the model. Without samples generated from VITA (i.e. *w/o gen*), the model performs poorly on various types of corruption. When the model is trained with samples generated from VITA and one type of vicinal samples (i.e. *w/o aug* or *w/o adv*), its corruption robustness is better than when training with both types of vicinal samples (i.e. *w/o gen*). However, its corruption robustness is still not as good as when vicinal samples are integrated from all sources. Besides, we can find that the combination of samples with shuffled adversarial perturbations and samples generated by VITA performs best (compared with *w/o adv* and *w/o gen*). This shows the superiority of the weakly augmented samples we generated compared to the common weakly augmented samples.

4.5 Robustness Towards Adversarial Attack

Adversarial Training Settings. Recent works (Rusak et al. 2020; Gilmer et al. 2019) substantiate the claim that increased robustness against regular or universal adversarial perturbations (Carlini and Wagner 2017; Moosavi-Dezfooli et al. 2017) does not imply increased robustness against common corruptions. In this part, we discuss whether our generated samples from VITA can be integrated into the existing adversarial training process to improve the adversarial robustness of the model. All images of CIFAR-10 are normalized into [0, 1]. The adversarial test data are bounded by l_∞ perturbations with $\epsilon_{test} = 0.031$. We use the same settings as the corruption robustness evaluation experiment to train our image-to-image translation framework. The only difference is the regularization terms in multi-source robust training. Here, we use strong adversarial examples via the regularization term proposed by TRADES (Zhang et al. 2019). Using a regularization term from TRADES (Zhang et al. 2019) and an early stopping scheme from FAT (Zhang et al. 2020), we deploy multi-source adversarial training to verify the effectiveness of samples generated by VITA towards adversarial robustness. The backbone of our network is Wide ResNet. Models are trained using SGD with 0.9 momenta for 100 epochs, with the initial learning rate of 0.01 divided by ten at epoch 60. Our supplementary materials contain further information about training and evaluating with additional models.

Adversarial Robustness Results. In Table 3, we can see that injecting the data generated by our method clearly improves the model's adversarial robustness against various adversarial attack methods. The usefulness of our suggested framework in strengthening model robustness against FGSM attack is particularly obvious. Although our VITA is not specifically designed for defending various adversarial attacks, we can easily increase adversarial robustness by simply altering the regularization terms.

5 Conclusion

In this work, we propose a multi-source vicinal transfer augmentation (VITA) method to mitigate performance degradation caused by off-manifold samples. The proposed VITA consists of two components: tangent transfer and integration of multi-source vicinal samples. To the best of our knowledge, our work is the first to reveal the effectiveness of tangents transfer for improving corruption robustness. Experimental results show that our proposed VITA obtains state-of-the-art performance on the image corruption benchmarks (CIFAR-10-C, CIFAR-100-C and ImageNet-C).

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant No. 61972188 and No. 62122035.

References

Azulay, A.; and Weiss, Y. 2019. Why do deep convolutional networks generalize so poorly to small image transformations? *J. Mach. Learn. Res.*, 20: 184:1–184:25.

Bengio, Y.; and Monperrus, M. 2004. Non-Local Manifold Tangent Learning. In *NIPS*, 129–136.

Bengio, Y.; Yao, L.; Alain, G.; and Vincent, P. 2013. Generalized Denoising Auto-Encoders as Generative Models. In *NIPS*, 899–907.

Carlini, N.; and Wagner, D. A. 2017. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy*, 39–57. IEEE Computer Society.

Chapelle, O.; Weston, J.; Bottou, L.; and Vapnik, V. 2000. Vicinal Risk Minimization. In *NIPS*, 416–422. MIT Press.

Chen, P.; Sharma, Y.; Zhang, H.; Yi, J.; and Hsieh, C. 2018. EAD: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples. In *AAAI*, 10–17. AAAI Press.

Cubuk, E. D.; Zoph, B.; Mané, D.; Vasudevan, V.; and Le, Q. V. 2019. AutoAugment: Learning Augmentation Strategies From Data. In *CVPR*, 113–123. Computer Vision Foundation / IEEE.

Dao, T.; Gu, A.; Ratner, A.; Smith, V.; Sa, C. D.; and Ré, C. 2019. A Kernel Theory of Modern Data Augmentation. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 1528–1537. PMLR.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Li, F. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*, 248–255. IEEE Computer Society.

Devries, T.; and Taylor, G. W. 2017. Improved Regularization of Convolutional Neural Networks with Cutout. *CoRR*, abs/1708.04552.

Dodge, S. F.; and Karam, L. J. 2017. A Study and Comparison of Human and Deep Learning Recognition Performance under Visual Distortions. In *ICCCN*, 1–7. IEEE.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting Adversarial Attacks With Momentum. In *CVPR*, 9185–9193. Computer Vision Foundation / IEEE Computer Society.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*. OpenReview.net.

Geirhos, R.; Temme, C. R. M.; Rauber, J.; Schütt, H. H.; Bethge, M.; and Wichmann, F. A. 2018. Generalisation in humans and deep neural networks. In *NeurIPS*, 7549–7561.

Gilmer, J.; Ford, N.; Carlini, N.; and Cubuk, E. D. 2019. Adversarial Examples Are a Natural Consequence of Test Error in Noise. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 2280–2289. PMLR.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative Adversarial Nets. In *NIPS*, 2672–2680.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR (Poster)*.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; Song, D.; Steinhardt, J.; and Gilmer, J. 2020a. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. *CoRR*, abs/2006.16241.

Hendrycks, D.; and Dietterich, T. G. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *ICLR (Poster)*. OpenReview.net.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2020b. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In *ICLR*. OpenReview.net.

Huang, G.; Liu, Z.; van der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *CVPR*, 2261–2269. IEEE Computer Society.

Isola, P.; Zhu, J.; Zhou, T.; and Efros, A. A. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *CVPR*, 5967–5976. IEEE Computer Society.

Kamann, C.; and Rother, C. 2020. Increasing the Robustness of Semantic Segmentation Models with Painting-by-Numbers. In *ECCV (10)*, volume 12355 of *Lecture Notes in Computer Science*, 369–387. Springer.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 1106–1114.

Lasserre, J. A.; Bishop, C. M.; and Minka, T. P. 2006. Principled Hybrids of Generative and Discriminative Models. In *CVPR (1)*, 87–94. IEEE Computer Society.

Lopes, R. G.; Smullin, S. J.; Cubuk, E. D.; and Dyer, E. 2020. Affinity and Diversity: Quantifying Mechanisms of Data Augmentation. *CoRR*, abs/2002.08973.

Lopes, R. G.; Yin, D.; Poole, B.; Gilmer, J.; and Cubuk, E. D. 2019. Improving Robustness Without Sacrificing Accuracy with Patch Gaussian Augmentation. *CoRR*, abs/1906.02611.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR (Poster)*. OpenReview.net.

Michaelis, C.; Mitzkus, B.; Geirhos, R.; Rusak, E.; Bringmann, O.; Ecker, A. S.; Bethge, M.; and Brendel, W. 2019. Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming. *CoRR*, abs/1907.07484.

Moosavi-Dezfooli, S.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal Adversarial Perturbations. In *CVPR*, 86–94. IEEE Computer Society.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2018. Do CIFAR-10 Classifiers Generalize to CIFAR-10? *CoRR*, abs/1806.00451.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI (3)*, volume 9351 of *Lecture Notes in Computer Science*, 234–241. Springer.

Rusak, E.; Schott, L.; Zimmermann, R. S.; Bitterwolf, J.; Bringmann, O.; Bethge, M.; and Brendel, W. 2020. A Simple Way to Make Neural Networks Robust Against Diverse Image Corruptions. In *ECCV (3)*, volume 12348 of *Lecture Notes in Computer Science*, 53–69. Springer.

Schneider, S.; Rusak, E.; Eck, L.; Bringmann, O.; Brendel, W.; and Bethge, M. 2020. Improving robustness against common corruptions by covariate shift adaptation. In *NeurIPS*.

Simard, P. Y.; LeCun, Y.; Denker, J. S.; and Victorri, B. 1996. Transformation Invariance in Pattern Recognition-Tangent Distance and Tangent Propagation. In *Neural Networks: Tricks of the Trade*, volume 1524 of *Lecture Notes in Computer Science*, 239–27. Springer.

Simard, P. Y.; Victorri, B.; LeCun, Y.; and Denker, J. S. 1991. Tangent Prop - A Formalism for Specifying Selected Invariances in an Adaptive Network. In *NIPS*, 895–903. Morgan Kaufmann.

Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. A. 2015. Striving for Simplicity: The All Convolutional Net. In *ICLR (Workshop)*.

Tang, Z.; Gao, Y.; Zhu, Y.; Zhang, Z.; Li, M.; and Metaxas, D. N. 2021. SelfNorm and CrossNorm for Out-of-Distribution Robustness. *CoRR*, abs/2102.02811.

Vasiljevic, I.; Chakrabarti, A.; and Shakhnarovich, G. 2016. Examining the Impact of Blur on Recognition by Convolutional Networks. *CoRR*, abs/1611.05760.

Xie, S.; Girshick, R. B.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *CVPR*, 5987–5995. IEEE Computer Society.

Yun, S.; Han, D.; Chun, S.; Oh, S. J.; Yoo, Y.; and Choe, J. 2019. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *ICCV*, 6022–6031. IEEE.

Zagoruyko, S.; and Komodakis, N. 2016. Wide Residual Networks. In *BMVC*. BMVA Press.

Zhang, H.; Cissé, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *ICLR (Poster)*. OpenReview.net.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E. P.; Ghaoui, L. E.; and Jordan, M. I. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 7472–7482. PMLR.

Zhang, J.; Xu, X.; Han, B.; Niu, G.; Cui, L.; Sugiyama, M.; and Kankanhalli, M. S. 2020. Attacks Which Do Not Kill Training Make Adversarial Learning Stronger. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, 11278–11287. PMLR.

Zhang, R. 2019. Making Convolutional Networks Shift-Invariant Again. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 7324–7334. PMLR.

Zhao, L.; Liu, T.; Peng, X.; and Metaxas, D. N. 2020. Maximum-Entropy Adversarial Data Augmentation for Improved Generalization and Robustness. In *NeurIPS*.

Zhu, J.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A. A.; Wang, O.; and Shechtman, E. 2017. Toward Multimodal Image-to-Image Translation. In *NIPS*, 465–476.