

PDE+: Enhancing Generalization via PDE with Adaptive Distributional Diffusion

Yige Yuan^{1,2}, Bingbing Xu^{1*}, Bo Lin³,

Liang Hou^{1,2}, Fei Sun¹, Huawei Shen^{1,2*}, Xueqi Cheng^{1,2*}

¹ CAS Key Laboratory of AI Security,

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

² University of Chinese Academy of Sciences

³ Department of Mathematics, National University of Singapore

{yuanyige20z,xubingbing,houliang17z,sunfei,shenhuawei,cxq}@ict.ac.cn, matbl@nus.edu.sg

Abstract

The generalization of neural networks is a central challenge in machine learning, especially concerning the performance under distributions that differ from training ones. Current methods, mainly based on the data-driven paradigm such as data augmentation, adversarial training, and noise injection, may encounter limited generalization due to model non-smoothness. In this paper, we propose to investigate generalization from a Partial Differential Equation (PDE) perspective, aiming to enhance it directly through the underlying function of neural networks, rather than focusing on adjusting input data. Specifically, we first establish the connection between neural network generalization and the smoothness of the solution to a specific PDE, namely “transport equation”. Building upon this, we propose a general framework that introduces adaptive distributional diffusion into transport equation to enhance the smoothness of its solution, thereby improving generalization. In the context of neural networks, we put this theoretical framework into practice as **PDE+** (**PDE** with **A**daptive **D**istributional **D**iffusion) which diffuses each sample into a distribution covering semantically similar inputs. This enables better coverage of potentially unobserved distributions in training, thus improving generalization beyond merely data-driven methods. The effectiveness of PDE+ is validated through extensive experimental settings, demonstrating its superior performance compared to SOTA methods.²

1 Introduction

The generalization of neural networks is a fundamental challenge in the field of machine learning. It refers to the ability of neural networks to perform effectively under unobserved distributions, which may differ from those encountered during the training process [7, 44]. Pursuing superior generalization capability is essential as it ensures model adaptability to diverse real-world scenarios, guaranteeing reliable predictions and decisions.

Existing approaches for improving generalization mainly employ a data-driven paradigm [18], including data augmentation [63], adversarial training [53], and noise injection [6]. In terms of implementation, they primarily enhance the training samples via manipulating the original input [75, 32, 53, 1] or transforming the hidden representations [70, 48, 51]. However, such a data-driven paradigm usually cannot guarantee reliable generalization capabilities on unobserved distributions. Taking data augmentation as an illustration, Fig. 1 shows that the model can only achieve satisfactory generalization performance when the training data is subjected to augmentation similar to that of the testing data. Analogous phenomena are also frequently observed in adversarial training and noise

*Corresponding author

²Code is available: <https://github.com/yuanyige/pde-add>.

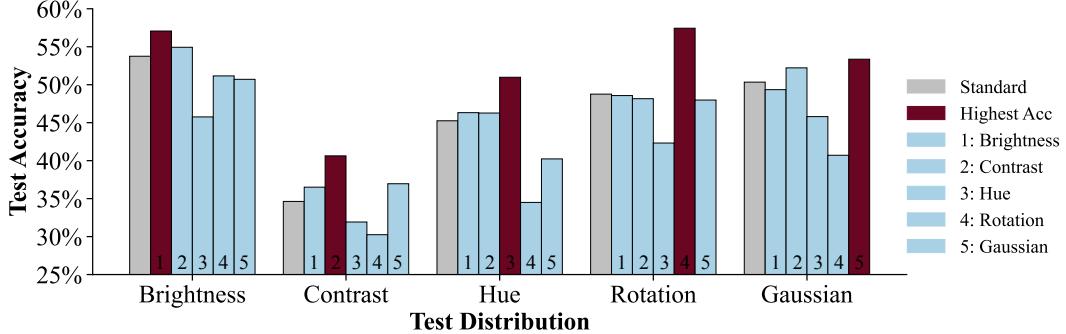


Figure 1: Model trained on six training distributions and evaluated on five corresponding test distributions. Model performs best on each test distribution is highlighted in red.

injection. For instance, adversarial training can improve generalization on adversarial examples but often comes with the cost of performance on natural data [67, 76]. Likewise, while injecting Gaussian noise can enhance generalization in the face of common corruptions, it risks σ -overfitting [37], i.e., overfitting to the particular Gaussian noise used in training.

The limited generalization capabilities of data-driven paradigm is due to model *irregularity* [71], i.e., the function learned by the neural network is non-smoothness. This may cause a problematic situation where semantically similar samples are encoded distantly, resulting in incorrect predictions. To address the irregularity issue, several approaches have been proposed to improve the smoothness of models [64, 66, 34], which helps to tackle the distribution shift problem [60]. Among them, Lipschitz continuity [64, 12] enforces smoothness constraints on models through regularization or architectural restrictions, e.g., gradient regularization [15] and spectral normalization [54]. However, such restrictions often come at the cost of expressive power [3].

In this paper, we go beyond the data-driven paradigm and propose to investigate generalization from a Partial Differential Equation (PDE) [5] perspective, aiming to directly introduce the smoothness constraint into the underlying function fn of neural network, rather than manipulating input data. The feasibility of this perspective is rooted in the intrinsic connection between neural networks and PDE [16, 47, 65, 28]. PDE describes a function that satisfies differential relationships, and neural networks can be regarded as a discrete numerical difference solver of PDE. That is to say, the underlying function of a neural network can be considered as the solution to PDE [47, 65, 28]. From such perspective, we can leverage the vast prior knowledge of PDE to constrain the underlying function of neural network, thus encouraging the resulting neural networks to exhibit specific desired properties, e.g., smoothness [71], well-posedness [26], and hyperbolicity [17].

The above fundamental connection inspires us to establish the connection between neural network generalization and the smoothness of PDE solution. Specifically, we initially model the neural network as the solution of a specific type of PDE, referred to as transport equation (TE) [47], which is often employed to describe the transportation of a quantity within a space. Then, a diffusion term is introduced into the TE, which has been proven to smooth the solution [71, 42]. The core of such paradigm is this key question: *What type of diffusion term is appropriate for a neural network to achieve effective generalization?* To answer it, we propose a general framework that introduces adaptive distributional diffusion into transport equation to enhance the smoothness of its solution. Such diffusion ensures suitable smoothness by treating the diffusion scope of each sample as a distribution that should cover the potential semantically similar inputs, thus improving generalization.

In the context of neural networks, we put this theoretical framework into practice as **PDE+** (**PDE** with **A**daptive **D**istributional **D**iffusion, PDE-ADD) to achieve generalization. Specifically, we introduce adaptive distributional diffusion into the neural network, which performs diffusion centered on each data point. The scope of each diffusion is modeled as a distribution, determined adaptively by multiple augmentations of the input. This enables better coverage of potentially unobserved distributions and improves generalization beyond data-driven approaches. The effectiveness of PDE+ is validated on various distributions, including clean samples and various common corruptions. The consistent improvements demonstrate the superior performance of our method over state-of-the-art methods.

Our main contributions include:

- (1) *A promising paradigm*: we investigate generalization from a Partial Differential Equation (PDE) perspective. To the best of our knowledge, we are the first to achieve generalization by establishing connections between the generalization of neural networks and the smoothness of TE solutions.
- (2) *An innovative method*: we propose an adaptive distributional diffusion term to incorporate smoothness into a neural network and instantiate it as PDE+, enabling better coverage of potentially unobserved distributions in training and improves generalization compared to data-driven methods.
- (3) *Solid experiments*: extensive experiments reveal PDE+ outperforms baselines across unobserved distributions, e.g., the improvements are up to 3.8% in Acc and 7.7% in mCE.

2 Related Work

In this section, we briefly review two lines of research that close to our work: the generalization of neural networks and differential equations based neural networks. Detailed introduction of related works can be found in Appendix B.

Generalization of Neural Networks. Current data-driven methods encompass data augmentation, adversarial training, and noise injection. Data augmentation is a widely adopted technique to enhance generalization, employing various strategies such as Mixup [75] and AugMix [32]. Adversarial training is a robust optimization approach for improving adversarial generalization [23] while potentially compromising non-adversarial generalization [67, 76]. Notable works in this area include PGD [53], TRADES [76], and RLAT [37]. Noise injection introduces noise into input data [2], activations [24], or hidden layers [10], whose noise magnitude can be sensitive and susceptible to overfitting [37]. Lipschitz continuity is often used to ensure model generalization [15, 54, 50], but its strict constraint can restrict a model’s capabilities [3]. Our method diverges from above approaches, as we directly constrain the smoothness of the neural network’s underlying function rather than fitting a finite set of input data like data-driven methods. Although our method shares the concept of smoothness with Lipschitz, it avoids compromising the model’s capabilities.

Differential Equations based Neural Networks The connection between continuous dynamical systems and residual neural networks [29] is initially established in [16]. Subsequently, numerous studies have delved into the relationships between various neural network architectures and different types of differential equations [52, 47, 65]. Since then, researchers have started to explore the beneficial properties of differential equations to enhance neural networks [72, 46, 71].

3 Generalization under PDEs with Adaptive Distributional Diffusion

This section introduces the theoretical motivation and framework behind our method. We begin by establishing connections between PDEs and neural networks, thereby transforming the generalization of neural networks into the smoothness of PDE solutions. Our innovative adaptive distributional diffusion term is then introduced to enhance the smoothness of solutions, which improves generalizability.

3.1 Neural Network as the Solution of Transport Equation

Partial Differential Equation (PDE) [5] is an equation containing an unknown function u of multiple variables and its partial derivatives. The connection between PDEs and neural networks has been discussed in [16], where neural networks could be interpreted as a numerical scheme to solve PDEs. Such connection allows us to take advantage of PDE, such as the properties of solution as well as the numerical schemes, to obtain a better neural network. In this section, we make use of the transport equation (TE), which is one special form of PDE, to interpret neural networks.

TE describes the concentration of a quantity transport in a fluid [57, 55] (Eq. (1)), which is suitable to model the feature transformation of data flow. This observation has also been discussed in [47, 65]

$$\frac{\partial u}{\partial t}(\mathbf{x}, t) + F(\mathbf{x}, \theta(t)) \cdot \nabla u(\mathbf{x}, t) = 0 \quad (1)$$

where $u(\mathbf{x}, t)$ denotes a function of concentration, which can be viewed as the underlying function of a neural network. $t \in (0, 1)$ denotes time, serving as the continuation of network layers. $\mathbf{x} \in \mathbb{R}^d$ denotes a variable in space, serving as the variable for data representation in terms of neural networks. ∇ represents gradient, and $F(\mathbf{x}, \theta(t))$ is the velocity field, serving as the continuation for network structures and parameters. In terms of neural networks, the changing of representation through layers can be viewed as a transport process over time. The representation is transported through each layer, where the parameters of each layer serve as a velocity field aiming to make changes to the sample representations and transport it to the next layer. Given the parameters of all layers, the representation transforms from the original input to final output, acting like a transport of data flow as illustrated in the top subfigure of Fig. 3.

$u(\mathbf{x}, t)$ represents the value obtained by transporting the variable \mathbf{x} through a series of $F(\mathbf{x}, \theta(t))$ from time t until the terminal. The terminal condition of TE is enforced at $t = 1$ as $u(\mathbf{x}, 1) = o(\mathbf{x})$, where $o(\mathbf{x})$ denotes the output function such as softmax [21]. Let $\hat{\mathbf{x}}$ denote the input feature. The original data-label pair $(\hat{\mathbf{x}}, y)$ is given at $t = 0$, and an optimal network u^* should exactly map $\hat{\mathbf{x}}$ to y , i.e., $u^*(\hat{\mathbf{x}}, 0) = y$. Obtaining the network is equivalent to solving the numerical solution of TE at $t = 0$ as $u(\hat{\mathbf{x}}, 0)$, where the method of characteristics [62] can be effectively employed. The main idea of the characteristics is to solve PDE via an ordinary differential equation (ODE) defining the characteristic curves of original PDE, which is shown in Eq. (2). Then the solution of PDE can be acquired by following these curves in Eq. (3).

$$d\mathbf{x}(t) = F(\mathbf{x}(t), \theta(t)) dt \quad (2)$$

$$u(\hat{\mathbf{x}}, 0) = o\left(\hat{\mathbf{x}} + \int_0^1 F(\mathbf{x}(t), \theta(t)) dt\right) \quad (3)$$

To solve Eq. (2) numerically, we adopt Euler method [9, Chapter 2] as shown in Eq. (4), which recovers the formulation of ResNet [29]. $l \in \{1, \dots, L\}$ is the network layer index, serving as a discrete slicing to continuous time t . \mathbf{h}_l and θ_l are representations and parameters at layer l , respectively.

$$\mathbf{h}_{l+1} = f(\mathbf{h}_l, \theta_l) + \mathbf{h}_l \quad (4)$$

$$u(\hat{\mathbf{x}}, 0) = o\left(\hat{\mathbf{x}} + \sum_{l=1}^L f(\mathbf{h}_l, \theta_l)\right) \quad (5)$$

Overall, neural network, particularly ResNet can be seen as a solution to TE. This connection lays a solid foundation to achieve desired properties of neural networks by constraining the solution of TE.

3.2 Improving Generalization via Enhancing the Smoothness of TE Solution

Smoothness has been demonstrated to be strongly linked to generalization, as it facilitates models to generalize beyond the training distribution [61, 60], enhances model robustness against small perturbations [12, 64], and plays a significant role in generalization quantization [34, 56] as well as uncertainty estimation [69, 49]. Building upon the insights, we propose to achieve generalization from the perspective of PDEs by modeling neural networks as solutions to PDEs and transforming the generalization goal of neural networks into smoothness goal of a solution to PDEs.

To enhance the smoothness of solution $u(\mathbf{x}, t)$, we leverage knowledge from PDE field to introduce a diffusion term [42] $\Delta u(\mathbf{x}, t)$ into TE as Eq. (6). The diffusion term corresponds to the Laplacian, i.e., the second-order derivative with respect to $\mathbf{x} \in \mathbb{R}^d$, as illustrated in Eq. (7). Here, Δ denotes the Laplacian operator, and $\sigma \neq 0$ is a coefficient for the diffusive magnitude.

$$\frac{\partial u}{\partial t}(\mathbf{x}, t) + F(\mathbf{x}, \theta(t)) \cdot \nabla u(\mathbf{x}, t) + \frac{1}{2} \sigma^2 \cdot \Delta u(\mathbf{x}, t) = 0 \quad (6)$$

$$\Delta u = \partial^2 u / \partial x_1^2 + \partial^2 u / \partial x_2^2 + \dots + \partial^2 u / \partial x_d^2 \quad (7)$$

Theorem 1 (Proved in Appendix C.1) *Given TE with diffusion term (Eq. (6)) with terminal condition $u(\mathbf{x}, 1) = o(\mathbf{x})$, where $F(\mathbf{x}, \theta(t))$ be a Lipschitz function in both \mathbf{x} and t , $o(\mathbf{x})$ be a bounded function. Then, for any small δ , $|u(\mathbf{x} + \delta, 0) - u(\mathbf{x}, 0)| \leq C \left(\frac{\|\delta\|_2}{\sigma} \right)^\alpha$ holds for constant $\alpha > 0$ if $\sigma \leq 1$, where $\|\delta\|_2$ is the ℓ_2 norm of δ , and C is a constant that depends on d , $\|o\|_\infty$, and $\|F\|_{L^\infty_{\mathbf{x}, t}}$.*

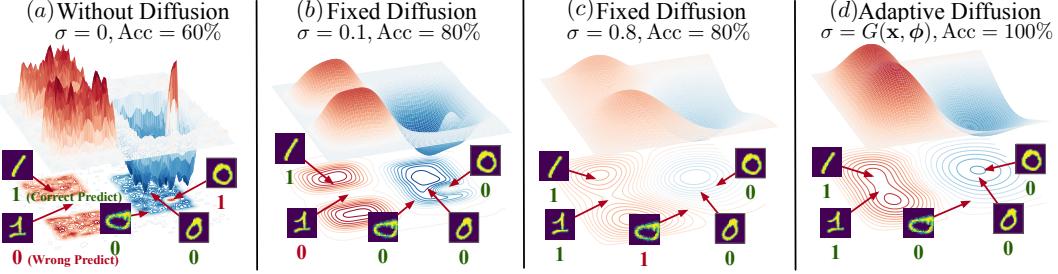


Figure 2: Solutions to 2D TE differs in the diffusion σ . The upper displays function surface, the lower exhibits its contour, with samples showing its true and predicted label.

Corollary 1 (Proved in Appendix C.2) *Generalization Error (GE) of model $u(\mathbf{x}, 0)$ trained on training set s_N is upper bounded by diffusion σ . For any $\epsilon > 0$, the following inequality holds with probability at least $1 - \epsilon$. For more details about the notations used, please refer to Appendix C.2.*

$$\text{GE}(u(\mathbf{x}, 0), s_N) \leq C \cdot L \left(\frac{\|\delta'\|_2}{\sigma} \right)^\alpha + M \sqrt{\frac{2K \ln 2 + 2 \ln(1/\epsilon)}{N}} \quad (8)$$

Typically, σ is chosen as a fixed scalar, imposing an uniform diffusion scale across entire data space [71]. Fixed diffusion brings smoothness into TE solution, but it neglects structure of solution for different \mathbf{x} . It cannot achieve an optimal diffusion scale for network across data space, as different locations require diverse diffusion scales based on their distance to other samples or class boundaries. To intuitively introduce the influence of diffusion, we illustrate the solution surface of 2D transport equation under different diffusion terms in Fig. 2. No diffusion in (a) results in highly irregular surface. Fixed diffusion with a small coefficient in (b) imposes insufficient smoothness for same class. Larger coefficient in (c) imposes over-smoothness for different classes. It reveals that a fixed coefficient can result in over-smoothness which diminishes variability, or insufficient smoothing. Thus, a new diffusion term is required to improve generalizability.

3.3 Adaptive Distributional Diffusion for Generalization

With concerns draw above, a crucial question arises:

What type of diffusion term is appropriate for a neural network to achieve effective generalization?

To address it, we claim that a good diffusion term for generalization should satisfy two goals: “Adaptive” and “Distributional”. “Adaptive” stands for that the diffusion varies in magnitude for every point across the entire space. “Distributional” treats the diffusion scope of each point as a distribution. For any input from the data space at any time step, the distribution should only encompass the inputs that are potentially similar to the central point in semantics. This mechanism allows for better coverage of potential unseen distributions and improved generalization compared to data-driven methods.

To achieve the above goals, we propose an **Adaptive Distributional Diffusion (ADD)** term and introduce it into TE as presented in Eq. (9). Rather than using a fixed scalar, our term incorporates a coefficient function $G(\mathbf{x}, \phi(t))$ that takes sample \mathbf{x} as input and outputs its diffusion scale, exhibiting different diffusion properties, based on the parameters ϕ at each time step t . The benefits of the term can be illustrated in Fig. 2(d), which allows for different smoothing effects across space in accordance with the principle of “adaptive”. Meanwhile, data spaces with similar semantics or within the same class can achieve smoothness in their scope, and those within different classes can avoid over-smoothness and maintain discrepancy. These satisfy the principle of “distributional”.

$$\frac{\partial u}{\partial t}(\mathbf{x}, t) + F(\mathbf{x}, \theta(t)) \cdot \nabla u(\mathbf{x}, t) + \frac{1}{2} G(\mathbf{x}, \phi(t))^2 \cdot \Delta u(\mathbf{x}, t) = 0 \quad (9)$$

3.4 Deriving Neural Network from Transport Equation with ADD

Introducing adaptive distributional diffusion into TE as Eq. (9) can realize the smoothness of the solution of TE, and thus encourage the resulting neural networks to exhibit generalization. In the following, we solve TE with ADD (Eq. (9)) to derive its corresponding neural network.

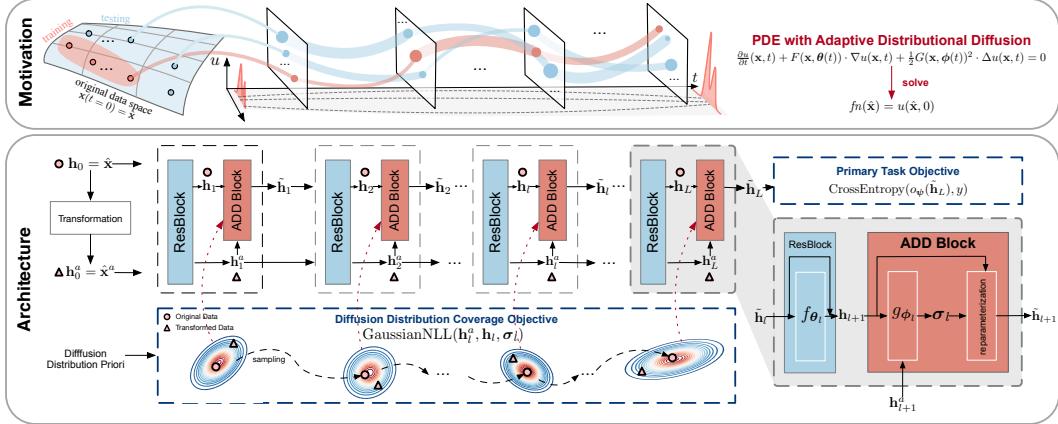


Figure 3: Motivation and architecture of PDE+, the upper illustrates our motivation of solving transport equation with adaptive distributional diffusion (ADD) to derive the functional form of neural network. The lower is the neural network instantiation, which comprises a series of blocks that contain a residual block followed by an ADD block. The architecture of the ADD block is enclosed grey frame on the right. The learning objectives are enclosed in two blue frames.

Theorem 2 (Proved in Appendix C.3) *TE with adaptive distributional diffusion term (Eq. (9)) can be solved using the Feynman-Kac formula [35]. The result is shown in Eqs. (10) and (11), where B_t represents the Brownian motion [68].*

$$u(\hat{\mathbf{x}}, 0) = \mathbb{E}[o(\mathbf{x}(1)) \mid \mathbf{x}(0) = \hat{\mathbf{x}}] \quad (10)$$

$$d\mathbf{x}(t) = F(\mathbf{x}(t), \theta(t)) dt + G(\mathbf{x}(t), \phi(t)) \cdot dB_t \quad (11)$$

The result is a conditional expectation with respect to the initial value problem of stochastic differential equation (SDE, [38]) in Eq. (11). To obtain the final functional form of our neural network, we adopt the Euler–Maruyama method [20] to compute the solution of SDE numerically as follows.

$$\begin{aligned} u(\hat{\mathbf{x}}, 0) &= \mathbb{E}[o(\mathbf{h}_L) \mid \mathbf{h}_0 = \hat{\mathbf{x}}] \\ \mathbf{h}_{l+1} &= \mathbf{h}_l + f(\mathbf{h}_l, \theta_l) + g(\mathbf{h}_l, \phi_l) \cdot \mathcal{N}(\mathbf{0}, \mathbf{I}) \end{aligned} \quad (12)$$

4 PDE+ : An Neural Network Instantiation

This section is for the instantiation of our framework **PDE+:** **PDE with Adaptive Distributional Diffusion** (PDE-ADD).

4.1 Overall Architecture

PDE+ is a neural network instantiation of PDE solution formulated in Eq. (12), where $\mathbf{h}_{l+1} = \mathbf{h}_l + f(\mathbf{h}_l, \theta_l)$ is the formulation for residual block, and $g(\mathbf{h}_l, \phi_l) \cdot \mathcal{N}(\mathbf{0}, \mathbf{I})$ is implemented as our adaptive distributional diffusion block, dubbed as ADD block. As shown in Fig. 3, the residual block is denoted as f_{θ_l} parameterized by θ_l , where $l \in \{1, \dots, L\}$ denotes the block index. ADD block is denoted as g_{ϕ_l} parameterized by ϕ_l . The overall architecture of PDE+, denoted as $f_{\theta, \phi}$ is the composition of L blocks, where each block contains a residual block followed by our ADD block.

4.2 Adaptive Distributional Diffusion Block

Fig. 3 illustrates the structure of ADD block, which takes the output from residual block \mathbf{h}_l as input, and outputs the scale σ_l for diffusion (Eq. (13)). Then a reparameterization trick [36] of \mathbf{h}_l and σ_l under the prior of Gaussian distribution is conducted to obtain the final output $\tilde{\mathbf{h}}_l$ (Eq. (14)).

$$\sigma_l = g_{\phi_l}(\mathbf{h}_l) \quad (13)$$

$$\tilde{\mathbf{h}}_l = \mathbf{h}_l + \sigma_l \cdot \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (14)$$

As introduced in Section 3.3, the principle of ADD blocks is “adaptive” and “distributional”. “Adaptive” is implemented by replacing the fixed diffusion with the learnable σ_l . “Distributional” means

that for any input from the data space at any given time step, the diffusion scope should encompass the potential neighbors that exhibit semantic similarity. To achieve this, semantically similar samples are utilized as guidance. Define training dataset s_N containing N training samples of C classes $s_N = \{(\mathbf{x}_n, y_n) \mid n \in 1, 2 \dots N\}$. Let \mathbf{x}_n^a represent samples that share semantic similarity with \mathbf{x}_n , such as augmented samples, style-transferred samples, or adversarial attack samples. We hope the diffusion distribution scope of \mathbf{x}_n can cover \mathbf{x}_n^a .

To achieve this, we let the original samples pass through the whole block with both residual block and ADD block, and the semantically similar samples only go through residual block without diffusion. Denote I as the identity function where $I(x) = x$. The l -th layer's representation of original samples and their semantically similar counterparts can be formulated as... Eqs. (15) and (16), .

$$\tilde{\mathbf{h}}_l = (g_{\phi_l} \circ (f_{\theta_{l-1}} + I) \circ \dots \circ g_{\phi_2} \circ (f_{\theta_1} + I))(\mathbf{x}) \quad (15)$$

$$\mathbf{h}_l^a = ((f_{\theta_{l-1}} + I) \circ \dots \circ (f_{\theta_1} + I))(\mathbf{x}^a) \quad (16)$$

For every block, the diffused hidden representation $\tilde{\mathbf{h}}_l$ can be regarded as a sampling from a Gaussian distribution $\mathcal{N}(\mathbf{h}_l, \sigma_l)$, where the representations of semantically similar samples \mathbf{h}_l^a should be covered. This objective can be implemented via maximizing the probability of \mathbf{h}_l^a under $\mathcal{N}(\mathbf{h}_l, \sigma_l)$ denoted as $p_\phi(\mathbf{h}_l^a \mid \mathbf{h}_l)$, which is equivalent to minimizing its negative log-likelihood. We named such objective as *diffusion distribution coverage objective* shown in Eq. (17), guiding only the parameters of diffusion blocks ϕ .

$$\min_{\phi} \mathbb{E}_{\mathbf{x} \sim s_N} - \sum_{l=1}^L \log p_{\phi_l}(\mathbf{h}_l^a \mid \mathbf{h}_l) = -\frac{1}{2N} \sum_{n=1}^N \sum_{l=1}^L \left[\log g_{\phi_l}(\mathbf{h}_l) + \frac{(\mathbf{h}_{n,l}^a - \mathbf{h}_{n,l})^2}{g_{\phi_l}(\mathbf{h}_l)} \right] \quad (17)$$

From a distributional point of view, the intuitive interpretation of our adaptive distributional diffusion is treating each sample as one distribution whose scope includes its semantic similar samples. Under such view, the basic residual block without diffusion treats each sample as a Dirac distribution [13] and our ADD block transforms it into Gaussian distribution. To broaden the distribution and enhance generalization, we advance from a single Gaussian to a Gaussian mixture [59], as it is a universal approximator of densities [22]. Notably , we do not model the Gaussian mixture distribution directly. Rather, we allow both the original sample and its augmentations to diffuse simultaneously, effectively acting as different Gaussian centers. As a result, the superimposition of these multiple single Gaussians manifests as a mixed Gaussian from a macroscopic perspective. This implementation can be easily achieved in one line of code, as shown in Algorithm 1(Algorithm 1) from Appendix D.

4.3 Learning Objectives

PDE+ consists of two learning objectives: a diffusion distribution coverage objective for every ADD block (Eq. (17)) and a primary task objective for the entire network. The primary task objective ensures the correctness of learning representations under diffusion. Define the output of \mathbf{x}_n throughout the whole model f_n as $\tilde{\mathbf{h}}_{n,L} = f_{\theta,\phi}(\mathbf{x}_n)$. The primary task objective is shown in Eq. (18), where o_ψ stands for output layer parameterized by ψ . The samples diffused throughout f_n to obtain a classification probability via softmax, guiding the learning of all parameters, including residual blocks θ , diffusion blocks ϕ and output ψ via cross-entropy. The algorithmic pseudocode for both training and testing phase can be found in Algorithms 1 and 2 in Appendix D.

$$\min_{\theta, \phi, \psi} \mathbb{E}_{(\mathbf{x}, y) \sim s_N} - \log p_{\theta, \phi, \psi}(y \mid \mathbf{x}) = -\frac{1}{N} \sum_{n=1}^N \left[\log \frac{\exp(o_\psi(\tilde{\mathbf{h}}_{n,L})_{y_n})}{\sum_{c=1}^C \exp(o_\psi(\tilde{\mathbf{h}}_{n,L})_c)} \right]_{y_n} \quad (18)$$

5 Experiments

In this section, we empirically evaluate PDE+ through the following questions. Due to the space limitations, more comprehensive experiments including full results on corruptions and diffusion scale analysis are provided in Appendix E.

- (Q1) Does PDE+ improve generalization compared to SOTA methods on various benchmarks?
- (Q2) Does PDE+ learn appropriate diffusion distribution coverage?
- (Q3) Does PDE+ improve generalization beyond observed (training) distributions?

Table 1: Comparisons of PDE+ and baselines on CIFAR-10(C), CIFAR-100(C) and Tiny ImageNet(C) based on ResNet-18. The corruption is evaluated under all severity level and the severest level. The best result is highlighted in **boldface**. The abbreviations means Standard (Std), Lipschitz (Lip), Noise Injection (NI), Data Augmentation (DA), Adversarial Training (AT).

Method	CIFAR-10(C)					CIFAR-100(C)					Tiny-ImageNet(C)				
	Clean		Corr Severity All		Corr Severity 5	Clean		Corr Severity All		Corr Severity 5	Clean		Corr Severity All		Corr Severity 5
	Acc (\uparrow)	Acc (\uparrow) mCE (\downarrow)	Acc (\uparrow)	mCE (\downarrow)	Acc (\uparrow)	Acc (\uparrow)	mCE (\downarrow)	Acc (\uparrow)	mCE (\downarrow)	Acc (\uparrow)	Acc (\uparrow)	mCE (\downarrow)	Acc (\uparrow)	mCE (\downarrow)	Acc (\uparrow) mCE (\downarrow)
Std ERM	95.35	74.63	100.00	57.19	100.00	77.71	49.27	100.00	33.18	100.00	54.02	25.57	100.00	15.54	100.00
Lip GradReg	93.64	77.62	96.29	62.33	91.52	73.80	52.16	96.95	37.33	94.49	52.01	29.20	95.13	19.91	94.86
NI EnResNet	83.33	74.34	137.98	66.87	63.72	67.11	49.28	103.61	40.24	83.56	49.26	25.83	100.18	19.01	96.55
NI RSE	95.59	77.86	94.12	63.66	89.08	77.98	53.73	94.10	38.03	92.88	53.74	27.99	96.81	18.92	96.11
NI NFM*	95.40	83.30	-	-	-	79.40	59.70	-	-	-	-	-	-	-	-
DA Gaussian	92.50	80.46	100.03	68.08	87.22	71.87	54.24	98.34	41.77	89.81	48.89	32.92	90.48	24.57	89.56
DA Mixup*	95.80	80.40	-	-	-	79.70	54.20	-	-	-	-	-	-	-	-
DA DeepAug*	94.10	85.33	64.63	77.29	60.05	-	-	-	-	-	54.90	-	-	-	-
DA AutoAug	95.61	85.37	61.74	75.12	62.07	76.34	58.72	83.12	45.38	82.84	52.63	35.14	87.67	25.36	88.54
DA AugMix	95.26	86.24	60.44	76.06	59.96	77.11	61.93	77.51	48.99	77.52	52.82	37.74	84.06	28.66	84.69
AT PGD $_{\ell_\infty}$	93.52	82.17	86.53	70.10	78.20	71.78	55.03	93.49	42.04	88.17	49.94	32.54	90.65	23.47	90.63
AT PGD $_{\ell_2}$	93.91	83.07	81.06	70.97	75.17	72.50	56.09	91.65	42.82	87.33	51.08	33.46	89.37	24.00	89.92
AT RLAT	93.23	83.67	80.98	72.73	72.59	71.10	56.54	91.98	44.27	86.24	50.24	33.13	89.83	24.46	89.47
AT RLAT _{Augmix}	94.73	88.28	55.60	80.37	51.56	75.06	62.77	77.38	51.60	74.24	51.29	37.92	83.69	29.05	84.17
Ours PDE+	95.59	89.11	48.07	82.81	44.97	78.84	65.62	69.68	54.22	69.43	53.72	39.41	81.80	30.32	82.68

Table 2: Single source domain generalization comparisons of PDE+ and baselines on PACS datasets based on ResNet-18 [29]. The best result is highlighted in **boldface**.

Source Domain	Method	Target Domain				Avg
		Photo	Art	Cartoon	Sketch	
Photo	ERM	-	21.33	22.31	28.35	24.00
	Augmix	-	26.90	24.10	27.05	26.02
	PDE+	-	25.43	28.58	37.69	30.57
Art	ERM	47.54	-	34.51	34.48	38.85
	Augmix	51.37	-	42.06	36.75	43.40
	PDE+	53.11	-	43.90	41.28	46.10
Cartoon	ERM	43.59	29.78	-	33.87	35.75
	Augmix	45.74	30.81	-	37.31	37.96
	PDE+	48.68	33.00	-	40.01	40.57
Sketch	ERM	18.74	16.16	25.26	-	20.05
	Augmix	26.28	26.51	45.34	-	32.72
	PDE+	30.05	30.90	45.43	-	35.47

injection based methods, including EnResNet [71], RSE [51], NFM [48]; Data augmentation based methods, including Gaussian noise, Mixup [75], DeepAug [30], AutoAug [14] and AugMix [32]; Adversarial training based methods, including PGD [53] and RLAT [37]. **(3) Metrics:** Accuracy is adopted as the main evaluation metric. Especially, for various corrupted distributions, mCE [31] is adopted for two severity levels: severity across all levels and the severest level 5. More comprehensive results for other severity and metrics are in Appendix E. **(4) Others:** According to Section 4.2, the semantically similar samples in PDE+ are generated using AugMix [32], a widely adopted data augmentation strategy that combines 7 distinct types of augmentations. It is important to note that we avoid overlap between these augmentations and the test distributions for most experiments.

5.1 Q1: PDE+ Outperforms SOTA on Benchmarks

Table 1 illustrate the results of PDE+ on CIFAR10(C), CIFAR100(C) and Tiny ImageNet(C) compared to baselines. “**” indicates that we reuse the results from [19, 37]. “-” indicates that this setting was not included in the paper. On original datasets, PDE+ achieves better performance than ERM, indicating that our diffusion does not obtain o.o.d. generalization at the cost of damaging performance on the original training distribution. The test distributions in corrupted datasets are different from training ones, which can be used to verify the effectiveness of generalization. Compared to numerous baselines across multiple categories, PDE+ achieves the best performance with respect to Acc and mCE on corruptions at the severest level and across all levels. The improvements are up to 3.8% in Acc and 7.7% in mCE. Such significant improvements make PDE+ stand out from other approaches that struggle to consistently improve performance across both original and diverse shifted distributions.

Experiments Settings A brief introduction of datasets, baselines and metrics is provided here, details can be found in Appendix F. **(1) Datasets:** Our experiments primarily focus on two types of datasets: *(i)* The original and 15 shift corruption distributions provided by CIFAR-10(C), CIFAR-100(C) and Tiny-ImageNet(C) [40, 43, 31]. *(ii)* The PACS dataset [45] encompasses four different domains: photo, art, cartoon, and sketch. **(2) Baselines:** we consider the representative and SOTA methods as baselines: standard training; Lipschitz continuity based gradient regularization [15]; Noise

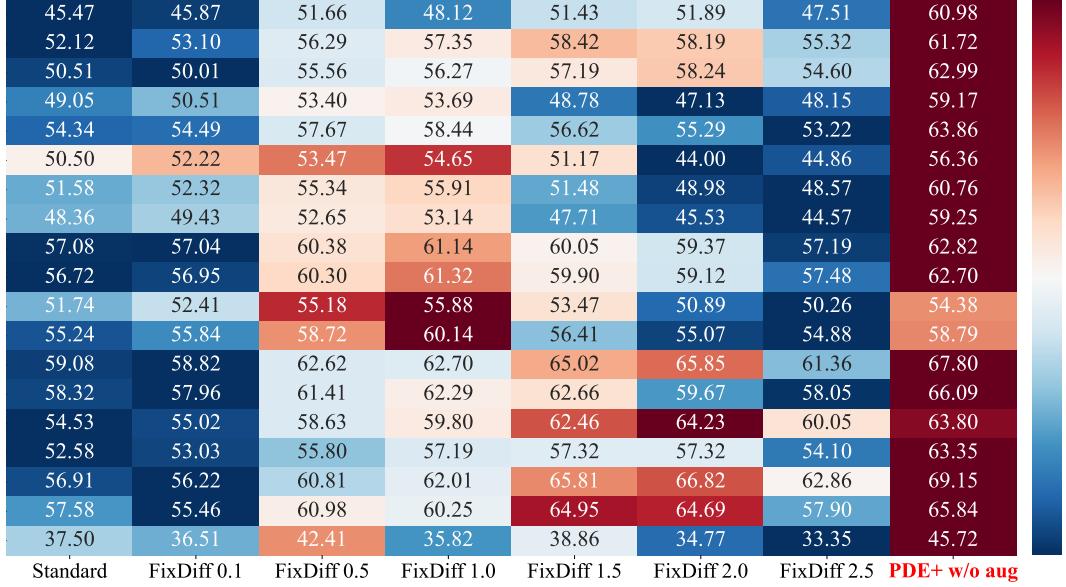


Figure 4: The heatmap of performance on neural networks with fixed diffusion (FixDiff) and PDE+ under fair comparison. The columns represent various training methods, with the FixDiff scale incrementally increasing from 0 to 2.5 across the first seven columns, while the last column is our PDE+. Each row corresponds to a unique test data distribution from the CIFAR-10-C dataset.

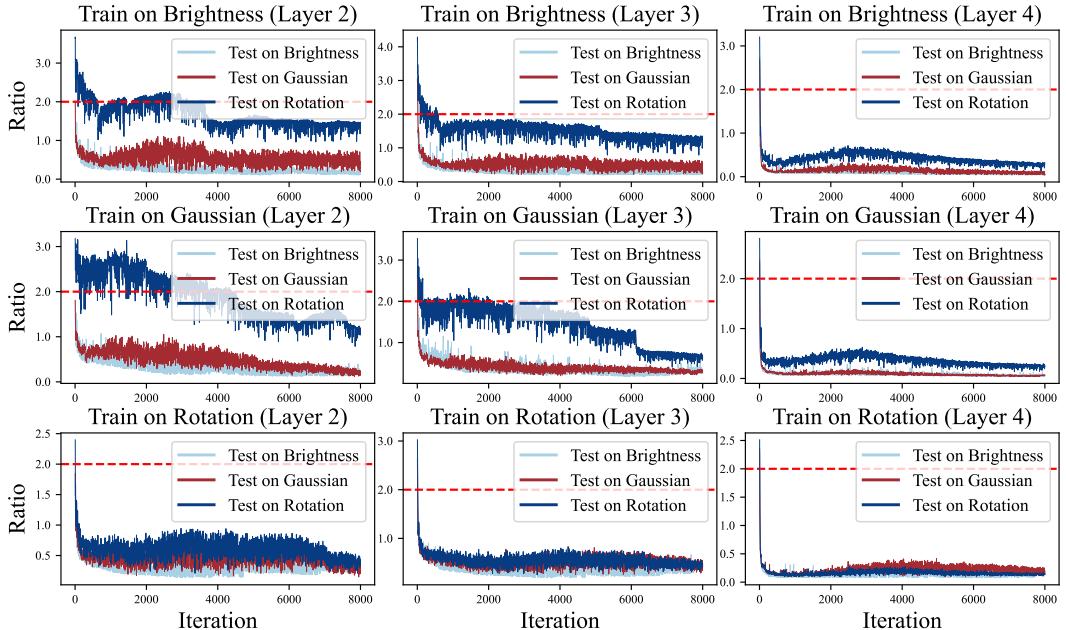


Figure 5: Diffusion coverage for unobserved distributions. Rows represent the training augmentations. Columns correspond to the layers of neural network. Each sub-figure includes three plots of distance- σ ratio during training for test samples generated by different test augmentations.

Table 2 illustrate results of PDE+ on PACS datasets. When training on a single source domain and testing on the remaining three domains, PDE+ surpasses baselines across all splits. This validates its efficacy not only in corrupted data, where distribution shifts may be relatively close, but also demonstrates effectiveness with cross-domain data where distribution shifts can be notably larger.

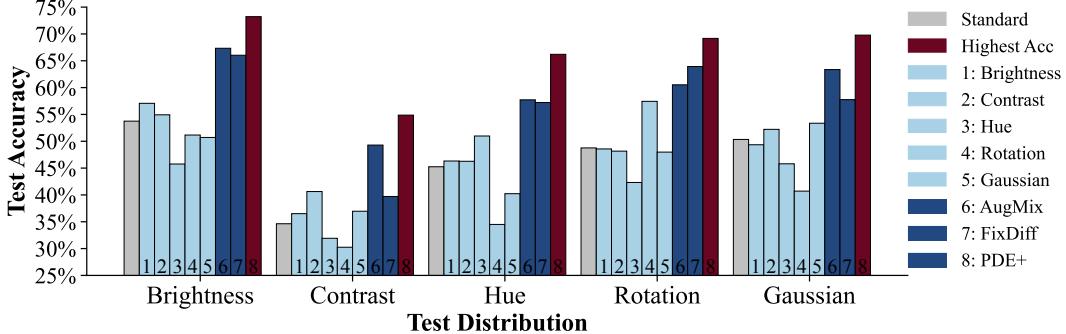


Figure 6: Generalization performance under five test distributions. The model is trained by different methods. The bar representing the highest performance is highlighted in red.

5.2 Q2: PDE+ Learns Appropriate Diffusion

This experiment is devoted to evaluating whether our proposed approach, whose diffusion scale is guided by augmented samples, can learn the appropriate diffusion scope. For a fair comparison, we do not conduct diffusion for augmented samples and only use augmented samples for the diffusion coverage guidance of the original samples (PDE+ w/o aug). This experiment can be viewed as the ablation study to evaluate if our learnable diffusion really works compared to fixed-scale diffusion (FixDiff for shorthand). Two conclusions can be drawn from the experimental results shown in Fig. 4: (1) Different corruption types, i.e., different distribution, prefers different magnitude/scale of smoothness, and a hard-to-please-everyone dilemma is caused by the fixed scale. (2) PDE+ indeed learns the appropriate diffusion scale. As is shown in the rightmost column, we can either achieve or be close to, the best performance of all corruption types.

5.3 Q3: PDE+ Generalizes Beyond Observation

This subsection aims to demonstrate that our method can generalize on distributions beyond training ones. Fig. 5 presents the changing trend of diffusion coverage for unobserved distributions, i.e., probabilities of unobserved test samples within the training diffusion distribution. This experiment is based on 2σ rule of Gaussian distribution, detailed description can be found in Appendix E.4. The results imply that even when training occurs on a single augmentation differing from testing, the likelihood of test samples being perceived as normal within the training diffusion distribution increases over time. Fig. 6 represent the experiment as an extension of the previous one on Fig. 1. Notably, PDE+ outperforms all other augmentations, including AugMix, demonstrating its capability of generalization on unobserved distributions.

6 Conclusion

In conclusion, we present a novel partial differential equations (PDE)-driven approach to address the generalization issue of neural networks across unseen data distributions, focusing on overcoming the limitations of data-driven methods. By modeling neural networks as solutions to PDEs in a transport equation framework, the connection between the solution smoothness of PDEs and the generalization of neural networks is established. The introduction of an adaptive distributional diffusion term helps improve the generalization of neural networks. An instantiation of this framework, called PDE+ can enhance the generalization via taking the augmented samples as semantic similar samples to guide the learning of adaptive distributional diffusion. Experimental results demonstrate the superior performance of PDE+ across various shifted distributions. This work opens up new avenues for research in generalization of neural networks from the PDE perspective and offers a promising direction for enhancing the generalization of neural networks.

Acknowledgments

This work was supported by the National KeyR&D Program of China (2022YFB3103700, 2022YFB3103704), the National Natural Science Foundation of China (NSFC) under Grants No.U21B2046 and No.62202448.

References

- [1] G. An. The effects of adding noise during backpropagation training on a generalization performance. *Neural Comput.*, 8(3):643–674, 1996.
- [2] G. An. The effects of adding noise during backpropagation training on a generalization performance. *Neural computation*, 8(3):643–674, 1996.
- [3] C. Anil, J. Lucas, and R. B. Grosse. Sorting out lipschitz function approximation, 2019.
- [4] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.
- [5] D. J. Arrigo. *An Introduction to Partial Differential Equations*. Synthesis Lectures on Mathematics & Statistics. Morgan & Claypool Publishers, 2017.
- [6] C. M. Bishop. Training with noise is equivalent to tikhonov regularization. *Neural Comput.*, 7(1):108–116, 1995.
- [7] O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.
- [8] H. Brezis and H. Brézis. *Functional analysis, Sobolev spaces and partial differential equations*, volume 2. Springer, 2011.
- [9] J. C. Butcher. *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons, 2003.
- [10] A. Camuto, M. Willetts, U. Simsekli, S. J. Roberts, and C. C. Holmes. Explicit regularisation in gaussian noise injections. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16603–16614. Curran Associates, Inc., 2020.
- [11] T. Chen, Z. Zhang, S. Liu, S. Chang, and Z. Wang. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*, 2021.
- [12] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, pages 854–863. PMLR, 2017.
- [13] S. B. Cohen and I. N. Kirschner. Approximating the dirac distribution for fourier analysis. *Journal of computational physics*, 93(2):312–324, 1991.
- [14] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [15] H. Drucker and Y. Le Cun. Improving generalization performance using double backpropagation. *IEEE Transactions on Neural Networks*, 3(6):991–997, 1992.
- [16] W. E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 1(5):1–11, 2017.
- [17] M. Eliasof, E. Haber, and E. Treister. Pde-gcn: novel architectures for graph neural networks motivated by partial differential equations. *Advances in neural information processing systems*, 34:3836–3849, 2021.
- [18] F. Emmert-Streib and M. Dehmer. Taxonomy of machine learning paradigms: A data-centric perspective. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(5):e1470, 2022.
- [19] N. B. Erichson, S. H. Lim, F. Utrera, W. Xu, Z. Cao, and M. W. Mahoney. Noisymix: Boosting robustness by combining data augmentations, stability training, and noise injections. *arXiv preprint arXiv:2202.01263*, 1, 2022.

- [20] M. Gelbrich and W. Römisch. Numerical solution of stochastic differential equations (peter e. kloeden and eckhard platen). *SIAM Rev.*, 37(2):272–275, 1995.
- [21] S. Gold, A. Rangarajan, et al. Softmax to softassign: Neural network algorithms for combinatorial optimization. *Journal of Artificial Neural Networks*, 2(4):381–399, 1996.
- [22] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.
- [23] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations*, 2015.
- [24] C. Gulcehre, M. Moczulski, M. Denil, and Y. Bengio. Noisy activation functions. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 3059–3068, New York, New York, USA, 20–22 Jun 2016. PMLR.
- [25] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.
- [26] E. Haber and L. Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34(1):014004, 2017.
- [27] W. W. Hager. Lipschitz continuity for constrained processes. *SIAM Journal on Control and Optimization*, 17(3):321–338, 1979.
- [28] J. Han, A. Jentzen, and W. E. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510, 2018.
- [29] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [30] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8320–8329, 2021.
- [31] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.
- [32] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan. Augmix: A simple method to improve robustness and uncertainty under data shift. In *International Conference on Learning Representations*, 2020.
- [33] T. Huster, C.-Y. J. Chiang, and R. Chadha. Limitations of the lipschitz constant as a defense against adversarial examples. In *ECML PKDD 2018 Workshops: Nemesis 2018, UrbReas 2018, SoGood 2018, IWAISe 2018, and Green Data Mining 2018, Dublin, Ireland, September 10-14, 2018, Proceedings 18*, pages 16–29. Springer, 2019.
- [34] P. Jin, L. Lu, Y. Tang, and G. E. Karniadakis. Quantifying the generalization error in deep learning in terms of data distribution and neural network smoothness. *Neural Networks*, 130:85–99, 2020.
- [35] M. Kac. On distributions of certain wiener functionals. *Transactions of the American Mathematical Society*, 65(1):1–13, 1949.
- [36] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [37] K. Klim, A. Maksym, and F. Nicolas. On the effectiveness of adversarial training against common corruptions. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.
- [38] P. E. Kloeden, E. Platen, P. E. Kloeden, and E. Platen. *Stochastic differential equations*. Springer, 1992.

- [39] S. Kotz, T. Kozubowski, and K. Podgórski. *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Number 183. Springer Science & Business Media, 2001.
- [40] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. *Citeseer*, 2009.
- [41] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
- [42] O. A. Ladyzhenskaia, V. A. Solonnikov, and N. N. Ural’tseva. *Linear and quasi-linear equations of parabolic type*, volume 23. American Mathematical Soc., 1968.
- [43] Y. Le and X. Yang. Tiny imangenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.
- [44] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [45] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.
- [46] M. Li, L. He, and Z. Lin. Implicit euler skip connections: Enhancing adversarial robustness via numerical stability. In *Proceedings of the 37th International Conference on Machine Learning, ICML’20*. JMLR.org, 2020.
- [47] Z. Li and Z. Shi. Deep residual learning and pdes on manifold. *CoRR*, abs/1708.05115, 2017.
- [48] S. H. Lim, N. B. Erichson, F. Utrera, W. Xu, and M. W. Mahoney. Noisy feature mixup. In *International Conference on Learning Representations*, 2022.
- [49] J. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax Weiss, and B. Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512, 2020.
- [50] W. Liu, J. Wang, H. Wang, R. Li, Y. Qiu, Y. Zhang, J. Han, and Y. Zou. Decoupled rationalization with asymmetric learning rates: A flexible lipshitz restraint. *arXiv preprint arXiv:2305.13599*, 2023.
- [51] X. Liu, M. Cheng, H. Zhang, and C.-J. Hsieh. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [52] Y. Lu, A. Zhong, Q. Li, and B. Dong. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net, 2018.
- [53] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.
- [54] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.
- [55] B. R. Munson, D. F. Young, T. H. Okiishi, and W. W. Huebsch. *Fundamentals of fluid mechanics*, john wiley & sons. Inc., USA, 2006.
- [56] N. Ng, N. Hulkund, K. Cho, and M. Ghassemi. Predicting out-of-domain generalization with local manifold smoothness. *CoRR*, abs/2207.02093, 2022.
- [57] N. Pogodaev. Optimal control of continuity equations. *Nonlinear Differential Equations and Applications NoDEA*, 23:1–24, 2016.
- [58] A. Raghunathan, S. M. Xie, F. Yang, J. Duchi, and P. Liang. Understanding and mitigating the tradeoff between robustness and accuracy. *arXiv preprint arXiv:2002.10716*, 2020.

- [59] D. A. Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.
- [60] P. Rodríguez, I. Laradji, A. Drouin, and A. Lacoste. Embedding propagation: Smoother manifold for few-shot classification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 121–138. Springer, 2020.
- [61] M. Rosca, T. Weber, A. Gretton, and S. Mohamed. A case for new neural network smoothness constraints. In J. Zosa Forde, F. Ruiz, M. F. Pradier, and A. Schein, editors, *Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops*, volume 137 of *Proceedings of Machine Learning Research*, pages 21–32. PMLR, 12 Dec 2020.
- [62] S. A. Sarra. The method of characteristics with applications to conservation laws. *Journal of Online mathematics and its Applications*, 3:1–16, 2003.
- [63] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *J. Big Data*, 6:60, 2019.
- [64] J. Sokolic, R. Giryes, G. Sapiro, and M. R. D. Rodrigues. Robust large margin deep neural networks. *IEEE Trans. Signal Process.*, 65(16):4265–4280, 2017.
- [65] Q. Sun, Y. Tao, and Q. Du. Stochastic training of residual networks: a differential equation viewpoint. *CoRR*, abs/1812.00174, 2018.
- [66] K. Than and N. Vu. Generalization of gans and overparameterized models under lipschitz continuity. *arXiv preprint arXiv:2104.02388*, 2021.
- [67] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.
- [68] G. E. Uhlenbeck and L. S. Ornstein. On the theory of the brownian motion. *Phys. Rev.*, 36:823–841, Sep 1930.
- [69] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*. PMLR, 2020.
- [70] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio. Manifold Mixup: Better Representations by Interpolating Hidden States. In *International Conference on Machine Learning (ICML)*, pages 6438–6447, 2019.
- [71] B. Wang, B. Yuan, Z. Shi, and S. J. Osher. Enresnet: Resnets ensemble via the feynman–kac formalism for adversarial defense and beyond. *SIAM Journal on Mathematics of Data Science*, 2(3):559–582, 2020.
- [72] Y.-J. Wang and C.-T. Lin. Runge-kutta neural network for identification of dynamical systems in high accuracy. *IEEE Transactions on Neural Networks*, 9(2):294–307, 1998.
- [73] H. Xu and S. Mannor. Robustness and generalization. In A. T. Kalai and M. Mohri, editors, *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pages 503–515. Omnipress, 2010.
- [74] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [75] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.
- [76] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. E. Ghaoui, and M. Jordan. Theoretically principled trade-off between robustness and accuracy. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR, 09–15 Jun 2019.

A Appendix Summary

The appendix contains the following sections:

- (1) Detailed related work (Appendix B).
- (2) Proofs:
 - Proof for Theorem 1 (Appendix C.1)
 - Proof for Corollary 1 (Appendix C.2)
 - Proof for Theorem 2 (Appendix C.3)
- (3) Algorithm (Appendix D).
- (4) Additional Experiments and Analysis:
 - Comprehensive results for each type of corruption (Appendix E.1)
 - Study on increasing severity (Appendix E.2)
 - Training curves, including diffusion coverage, loss, and accuracy (Appendix E.3)
 - Analysis for Diffusion Coverage of Unseen Distributions (Appendix E.4)
- (5) Settings:
 - The summary of datasets (Appendix F.1)
 - The baselines (Appendix F.2)
 - The metrics (Appendix F.3)
 - The hyper-parameters (Appendix F.4)
 - The computing resources (Appendix F.5)
- (6) The limitations and future explorations (Appendix G).

B Detailed Related Work

B.1 Data Augmentation

Data augmentation is a widely adopted technique for enhancing the generalization performance of machine learning models. By applying various transformations to input data, data augmentation effectively increases the size of the training dataset, encouraging models to learn more generalized features. A general form is shown in Eq. (19), where θ and L denote the parameter and loss, \mathcal{D} denotes the training dataset, (x, y) denotes the original data and its labels, x' denotes the augmented data. Classical data augmentation techniques include flipping, rotation, cropping, and color jittering. More recent methods have been proposed to further improve generalization. Mixup [75] interpolates between both feature and label of two samples; CutMix [74] combines segments of two images in a patch-wise manner; DeepAug [30] employs reinforcement learning to learn augmentation strategies from data; AugMix [32] mixes augmented images and enforces consistent embeddings of the augmented images, which serves as a strong data augmentation technique for generalization. Although data augmentation is an intuitive yet powerful method, it is limited by its data-driven nature, as the choice of specific augmentation strategies remains a limitation in terms of generalizaiotn across distributions. For example, a Gaussian augmentation strategy may struggle when applied to images with illumination change. Even with the combination of multiple augmentation strategies, the generalization is still unassured for distribution outside the combination.

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} [L(\theta, x', y)] \quad (19)$$

Our method PDE+, serving as a global smoothness constraint via the lens of PDE, can be directly understood as follows: Based on the smoothness assumption, our mechanism smooths the gap between original sample and augmented sample with random sampled noise. In other words, we create a more continuous and smoothed surrounding from the original sample to cover the augmented sample. PDE+ enables better coverage of potentially unobserved distributions within the gap or surrounding, thereby improving generalization beyond data augmentation.

B.2 Adversarial Training

Adversarial Training (AT) is a robust optimization technique designed to improve adversarial robustness, i.e., the robustness of machine learning models against adversarial examples, which are small but maliciously perturbed inputs that can deceive models into making incorrect predictions.

A general form is shown in Eq. (20), where δ and \mathcal{S} denote the perturbation and its radius. The concept was initially introduced by [23], and several AT approaches have been proposed over the years, including Fast Gradient Sign Method (FGSM) [23], Projected Gradient Descent (PGD) [53], and TRADES [76]. Although AT can significantly improve generalization on adversarial examples, two phenomena have been observed: (1) It exists an inherent trade-off between the generalization on adversarial examples and original clean samples, which has been widely observed in [67, 76]. (2) AT enforces models robust against adversarial attacks of a specific type and certain magnitudes, its gained robustness does not extrapolate to larger perturbations nor unseen attack types [11]. These observed phenomena underscore the limitations inherent in the data-driven nature of AT. In essence, the model merely fit, or even overfit, the training examples it has observed [58, 11]. This notion aligns with the proposition that the trade-off between standard and robust error is likely to diminish given an infinite dataset [58], or injecting more learned smoothness [11].

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\delta \in \mathcal{S}} L(\theta, x + \delta, y) \right] \quad (20)$$

While the primary objective of AT is typically to enhance adversarial robustness, our aims diverge to concentrate on a broader concept of generalization across diverse distributions. There exist works that exploit AT to bolster generalization, and it has been discovered that the judicious selection of perturbation radius during AT can enhance non-adversarial generalization on common corruptions [37]. Our method, PDE+, can be construed as a synergy of "smoothness" and "appropriate radius selection", thus sharing common insights from these aforementioned works.

B.3 Noise Injection

Noise Injection is another technique aiming at improving the generalizatio of machine learning models. It involves injecting noise into input data [2], activation functions [24], or hidden layers [10]. Several noise types have also been explored, including Gaussian noise [71], dropout noise [65], and Bernoulli noise. Recent works, such as [37, 48], have demonstrated that training models with noise can improve their generalization to unseen corruptions. Existing methods are as follows, EnResNet [71] improves performance through an ensemble of ResNets with injected noise. RSE [51] presents a framework that injects noise and employs self-ensembling during testing to enhance model robustness. NFM [48] combines noise injection and manifold mixup [70] to expand the generalization capacity of the model. However, the magnitude of noise plays a significant role in determining the effectiveness of noise injection. It has been shown in [37] that noise injection tends to overfit to a particular magnitude of noise used for training, resulting in a significant detrimental effect on generalization, which is identified as σ -overfitting. PDE+ conducts adaptive diffusion grounded in the principles of PDEs to ensure smoothness, avoiding overfitting to a particular magnitude.

B.4 Lipschitz Continuity

Lipschitz continuity (shown in Eq. (21)) is a mathematical property that serves as a popular constraint on the smoothness of functions, ensuring that the output of a function does not change too drastically with respect to small changes in its input [27]. Given two metric spaces (X, d_X) and (Y, d_Y) , where d_X denotes the metric on the set X and d_Y is the metric on set Y , a function $f : X \rightarrow Y$ is called Lipschitz continuous if there exists a real constant $K \geq 0$ such that, for all x_1 and x_2 in X ,

$$d_Y(f(x_1), f(x_2)) \leq K d_X(x_1, x_2) \quad (21)$$

In the field of deep learning, Lipschitz continuity has become a key aspect in improving model generalization. Methods grounded in Lipschitz continuity can be broadly categorized into two groups: architectural constraints and regularization approaches. Architectural constraints place limitations on the operator norm, e.g., weight clipping [4] controls the Lipschitz constant by setting bounds on the weights. Spectral normalization [54] normalizes the spectral norm of the model's weight matrices. These methods are proved to satisfy Lipschitz constraints. However, they may limit the expressive capacity of the model. For instance, it's been shown that norm-constrained ReLU networks cannot approximate simple functions such as the absolute value [33]. Regularization approaches impose smoothness by regularization. For instance, [15] imposes smoothness by applying regularization on gradients. [25] conducts gradient penalty to fix the poor gradient behaviors of weight clipping. Although these methods show promising results in practice, they do not guarantee Lipschitz constraint on a global scale [3].

C Proof

C.1 Proof for Theorem 1

Recalling that in Theorem 1 we have, Given TE with diffusion term (Eq. (6)) with terminal condition $u(\mathbf{x}, 1) = o(\mathbf{x})$, where $F(\mathbf{x}, \theta(t))$ be a Lipschitz function in both \mathbf{x} and t , $o(\mathbf{x})$ be a bounded function. Then, for any small δ , $|u(\mathbf{x} + \delta, 0) - u(\mathbf{x}, 0)| \leq C\left(\frac{\|\delta\|_2}{\sigma}\right)^\alpha$ holds for constant $\alpha > 0$ if $\sigma \leq 1$, where $\|\delta\|_2$ is the ℓ_2 norm of δ , and C is a constant that depends on d , $\|o\|_\infty$, and $\|F\|_{L_{\mathbf{x}, t}^\infty}$.

Proof 1 For simplicity, here we only illustrate the case for $d = 1$ in the initial value problem (i.e., $t \in (0, T)$ given $u(\mathbf{x}, 0)$), where the terminal value problem can be proved by reverse time. General proof and the details of following notations could be found in [42]. The outline of this proof is as follows: First, we adopt the Sobolev embedding theorem for embedding H^k space into Hölder spaces, which constructs the left-hand side of the main inequality in Theorem 1. Consequently, the remainder of the proof applies energy method to give an H^k estimate of solution u .

According to Sobolev embedding theorem [8], for $\alpha < k - \frac{d}{2}$, it holds that

$$\sup_{\mathbf{x}, \delta} \frac{\|u(\mathbf{x} + \delta, T) - u(\mathbf{x}, T)\|}{\|\delta\|_2^\alpha} \leq C_1 \|u(\mathbf{x}, T)\|_{H^k} \quad (22)$$

where constant C_1 depends on d and k . For $d = 1$, it suffices to prove Eq. (22) for $k = 1$. Multiplying Eq. (6) by u and integrating it for variable \mathbf{x} , we obtain

$$\begin{aligned} \frac{d}{dt} \|u\|_{L^2}^2 + 2\sigma \|\nabla u\|_{L^2}^2 &\leq 2\|F\|_{L_{\mathbf{x}, t}^\infty} \|\nabla u\|_{L^2} \|u\|_{L^2} \\ &\leq \sigma \|\nabla u\|_{L^2}^2 + \frac{\|F\|_{L_{\mathbf{x}, t}^\infty}^2 \|u\|_{L^2}^2}{\sigma} \end{aligned} \quad (23)$$

which implies

$$\frac{d}{dt} \|u\|_{L^2}^2 + \sigma \|\nabla u\|_{L^2}^2 \leq \frac{\|F\|_{L_{\mathbf{x}, t}^\infty}^2 \|u\|_{L^2}^2}{\sigma} \quad (24)$$

Then by Grönwall's inequality, $\|u\|_{L^2}$ satisfies $\|u(\mathbf{x}, t)\|_{L^2}^2 \leq e^{\|F\|_{L_{\mathbf{x}, t}^\infty}^2 t/\sigma} \|u(\mathbf{x}, 0)\|_{L^2}^2$. Plugging the estimate of $\|u\|_{L^2}$ back to Eq. (24), it is easy to see

$$\|\nabla u(\mathbf{x}, t)\|_{L^2}^2 \leq \frac{\|F\|_{L_{\mathbf{x}, t}^\infty}^2}{\sigma^2} e^{\|F\|_{L_{\mathbf{x}, t}^\infty}^2 t/\sigma} \|u(\mathbf{x}, 0)\|_{L^2}^2 \quad (25)$$

which suggests

$$\|u(\mathbf{x}, t)\|_{H^1}^2 \leq \frac{\|F\|_{L_{\mathbf{x}, t}^\infty}^2 + 1}{\sigma^2} e^{\|F\|_{L_{\mathbf{x}, t}^\infty}^2 t/\sigma} \|u(\mathbf{x}, 0)\|_{L^2}^2 \quad (26)$$

Plugging Eq. (26) into Eq. (22) yields the desired statement.

C.2 Proof for Corollary 1

Recalling that in Corollary 1 we have, Generalization Error (GE) of model $u(\mathbf{x}, 0)$ trained on training set s_N is upper bounded by diffusion σ . For any $\epsilon > 0$, the following inequality holds with probability at least $1 - \epsilon$.

$$\text{GE}(u(\mathbf{x}, 0), s_N) \leq C \cdot L \left(\frac{\|\delta'\|_2}{\sigma} \right)^\alpha + M \sqrt{\frac{2K \ln 2 + 2 \ln(1/\epsilon)}{N}}$$

Proof 2 Assume that the loss function is denoted as $l(y, f(x))$, which quantifies the discrepancy between the true label y and the predicted label by the model $f(x)$. A training set is denoted as s_N , where $\{(x_{s_i}, y_{s_i})\}_{i=1}^N \in s_N$ are N samples drawn from the entire data distribution \mathcal{D} .

Let \mathcal{A}_{s_N} be a learning algorithm trained on the training set s_N . The empirical risk of \mathcal{A}_{s_N} (Eq. (27)), while the expected risk of \mathcal{A}_{s_N} on the whole data distribution μ is defined in Eq. (28). Consequently,

the definition for generalization (generalization error [73]), which measures the difference between empirical risk and expected risk, is shown in Eq. (29).

$$\ell_{emp}(f, s_N) \triangleq \frac{1}{N} \sum_{i=1}^N \ell(y_{s_i}, f(x_{s_i})) \quad (27)$$

$$\ell_{exp}(f) \triangleq \mathbb{E}_{(x,y) \sim \mu} [\ell(y, f(x))] \quad (28)$$

$$GE(f, s_N) \triangleq |\ell_{exp}(f) - \ell_{emp}(f, s_N)| \quad (29)$$

In our case, $f(x)$ is equivalent to the solution of TE, i.e., $u(x, 0)$. Therefore, our generalization error is defined as:

$$GE(u(x, 0), s_N) \triangleq |\ell_{exp}(u(x, 0), y) - \ell_{emp}(u(x_s, 0), y_s)|$$

Referring back to Theorem 1, similar to [73], for a K -way classification problem, we partition the data space \mathcal{Z} into K partitions $\{P_i\}_{i=1}^K$, where P_i signifies the partition for the i th class. Let δ' denote the maximum potential perturbation for each partition, and l represent the L -Lipschitz loss function which is upper bounded by M . Therefore, for any x_s and x in the same class P_i , Eq. (31) holds.

$$|\ell(u(x, 0), y) - \ell(u(x_s, 0), y_s)| \quad (30)$$

$$\leq C \cdot L \left(\frac{\|x - x_s\|_2}{\sigma} \right)^\alpha \leq C \cdot L \left(\frac{\|\delta'\|_2}{\sigma} \right)^\alpha \quad (31)$$

We can derive a diffusion coefficient upper bound to generalization error. Let N_i be the set of training samples of s_N that fall into the P_i . Due to Breteganoille-Huber-Carol inequality[73], Eq. (32) holds with probability at least $1 - \epsilon$.

$$\sum_{i=1}^K \left| \frac{|N_i|}{n} - \mu(P_i) \right| \leq \sqrt{\frac{2K \ln 2 + 2 \ln(1/\epsilon)}{N}} \quad (32)$$

The derivation is as follows, where (a), (b), and (c) are due to the triangle inequality, the definition of N_i , and the smoothness bound from Theorem 1 of the original paper. (d) is due to Eq. (32), which connects generalization with diffusion.

$$\begin{aligned} GE(u(x, 0), s_N) &\triangleq |\ell_{exp}(u(x, 0), y) - \ell_{emp}(u(x_s, 0), y_s)| \\ &= \left| \sum_{i=1}^K \mathbb{E}(\ell(u(x, 0), y) | (x, y) \in P_i) \mu(P_i) - \frac{1}{N} \sum_{i=1}^N \ell(u(x_{s_i}, 0), y_{s_i}) \right| \\ &\stackrel{(a)}{\leq} \left| \sum_{i=1}^K \mathbb{E}(l(u(x, 0), y) | (x, y) \in P_i) \frac{|N_i|}{N} - \frac{1}{N} \sum_{i=1}^N l(u(x_{s_i}, 0), y_{s_i}) \right| \\ &\quad + \left| \sum_{i=1}^K \mathbb{E}(l(u(x, 0), y) | (x, y) \in P_i) \mu(P_i) - \sum_{i=1}^K \mathbb{E}(l(u(x, 0), y) | (x, y) \in P_i) \frac{|N_i|}{N} \right| \\ &\stackrel{(b)}{\leq} \left| \frac{1}{N} \sum_{i=1}^K \sum_{j \in N_i} \max_{(x,y) \in P_i} |l(u(x_{s_j}, 0), y_{s_j}) - l(u(x, 0), y)| \right| + \left| \max_{(x,y) \in \mathcal{Z}} |l(u(x, 0), y)| \sum_{i=1}^K \left| \frac{|N_i|}{N} - \mu(P_i) \right| \right| \\ &\stackrel{(c)}{\leq} C \cdot L \left(\frac{\|\delta'\|_2}{\sigma} \right)^\alpha + M \sum_{i=1}^K \left| \frac{|N_i|}{N} - \mu(P_i) \right| \stackrel{(d)}{\leq} C \cdot L \left(\frac{\|\delta'\|_2}{\sigma} \right)^\alpha + M \sqrt{\frac{2K \ln 2 + 2 \ln(1/\delta)}{N}} \end{aligned}$$

C.3 Proof for Theorem 2

Recalling that in Theorem 2 we have, TE with adaptive distributional diffusion term (Eq. (9))can be solved using the Feynman-Kac formula [35], The result is shown in Eqs. (10) and (11), where B_t represents the Brownian motion [68].

$$u(\hat{\mathbf{x}}, 0) = \mathbb{E}[o(\mathbf{x}(1)) | \mathbf{x}(0) = \hat{\mathbf{x}}] \quad (33)$$

$$d\mathbf{x}(t) = F(\mathbf{x}(t), \boldsymbol{\theta}(t)) dt + G(\mathbf{x}(t), \boldsymbol{\phi}(t)) \cdot dB_t \quad (34)$$

Proof 3 Assume that

$$d\mathbf{x}_t = F(\mathbf{x}_t, t)dt + G(\mathbf{x}_t, t)dB_t \quad (35)$$

$$\mathbf{x}_t = \mathbf{x}, t < T \quad (36)$$

Let $u(\mathbf{x}, t) = \mathbb{E}[o(\mathbf{x}_T) | \mathcal{F}_t]$. Eq. (35) shows \mathbf{x}_t is a Markov process, and therefore $u(\mathbf{x}, t) = \mathbb{E}[o(\mathbf{x}_T) | \mathbf{x}_t = \mathbf{x}]$. Moreover,

$$\begin{aligned} u(\mathbf{x}, t) &= \mathbb{E}[o(\mathbf{x}_T) | \mathcal{F}_t] \\ &= \mathbb{E}[\mathbb{E}[o(\mathbf{x}_T) | \mathcal{F}_{t+dt}] | \mathcal{F}_t] \\ &= \mathbb{E}[u(\mathbf{x}_{t+dt}, t+dt) | \mathcal{F}_t] \\ &= \mathbb{E}[u(\mathbf{x}_t, t) + \frac{\partial u}{\partial t} dt + \frac{\partial u}{\partial \mathbf{x}} d\mathbf{x}_t + \frac{1}{2} \frac{\partial^2 u}{\partial \mathbf{x}^2} d[\mathbf{x}, \mathbf{x}](t) | \mathcal{F}_t] \\ &= u(\mathbf{x}, t) + \frac{\partial u}{\partial t} dt + \frac{\partial u}{\partial \mathbf{x}} \mathbb{E}[d\mathbf{x}_t | \mathcal{F}_t] + \frac{1}{2} \frac{\partial^2 u}{\partial \mathbf{x}^2} \mathbb{E}[d[\mathbf{x}, \mathbf{x}](t) | \mathcal{F}_t] \end{aligned} \quad (37)$$

According to Eq. (35), the above equation yields

$$\frac{\partial u}{\partial t} + F(\mathbf{x}, t) \frac{\partial u}{\partial \mathbf{x}} + \frac{1}{2} G^2(\mathbf{x}, t) \frac{\partial^2 u}{\partial \mathbf{x}^2} = 0 \quad (38)$$

D Algorithm

The following pseudocodes provide an overview of PDE+. The training process is described in Algorithm 1, the testing process is described in Algorithm 2.

Algorithm 1: Training Phase of PDE+

Input: Training dataset (\mathbf{x}, y) . Number of blocks L . Data Augmentor

Output: Trained parameters θ, ϕ and ψ

while $epoch \leq MAX_ITER$ **do**

```

/* forwarding for every layer's optimal scale */
```

```

 $\mathbf{x}^a = \text{Augmentor}(\mathbf{x})$ 
```

```

 $\tilde{\mathbf{h}}_0 = \mathbf{x}$ 
```

```

 $\mathbf{h}_0^a = \mathbf{x}^a$ 
```

```

for  $l = 1, 2, \dots, L$  do
```

```

    /* do convection for original data and augmented data */
```

```

     $\mathbf{h}_l = f_\theta(\tilde{\mathbf{h}}_{l-1})$ 
```

```

     $\mathbf{h}_l^a = f_\theta(\mathbf{h}_{l-1}^a)$ 
```

```

    /* do diffusion only for original data */
```

```

     $\boldsymbol{\sigma}_l = g_{\phi_l}(\mathbf{h}_l)$ 
```

```

     $\tilde{\mathbf{h}}_l = \mathbf{h}_l + \boldsymbol{\sigma}_l \cdot \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
```

```

    compute loss  $- \log p(\mathbf{h}_l^a | \mathbf{h}_l)$  for every layers according to Eq. (17) to update  $\phi$ .
```

```

/* forwarding for primary task */
```

```

 $\mathbf{x} = \text{concat}(\mathbf{x}, \mathbf{x}^a)$ 
```

```

 $\tilde{\mathbf{h}}_L = u_{\theta, \phi}(\mathbf{x})$ 
```

```

compute loss  $- \log p(y | \mathbf{x})$  for the last layer according to Eq. (18) to update  $\theta, \phi$  and  $\psi$ .
```

E Experiments and Analysis

This section delves into an expanded set of experiments and associated analysis. These include (1) Comprehensive results for each type of corruption, as detailed in Appendix E.1; (2) Exploration of increasing severity impact, which can be found in Appendix E.2; and (3) An examination of training trends, specifically those concerning diffusion coverage, loss, and accuracy, which are elucidated in Appendix E.3.

Algorithm 2: Testing Phase of PDE+

Input: Testing dataset \mathbf{x} . Ensemble iter E
Output: Prediction \hat{y}

```

1 initialize  $\mathbf{h} = (0, 0, \dots, 0)$ 
  /* do ensemble for multiple times */
2 for  $i = 1, 2 \dots E$  do
3   |  $\tilde{\mathbf{h}}_L = u_{\theta, \phi}(\mathbf{x})$ 
4   |  $\mathbf{h} = \mathbf{h} + \tilde{\mathbf{h}}_L$ 
5 end
  /* predict label according to the max ensembling probability */
6  $\hat{y} = \arg \max o_{\phi}(\mathbf{h})$ 

```

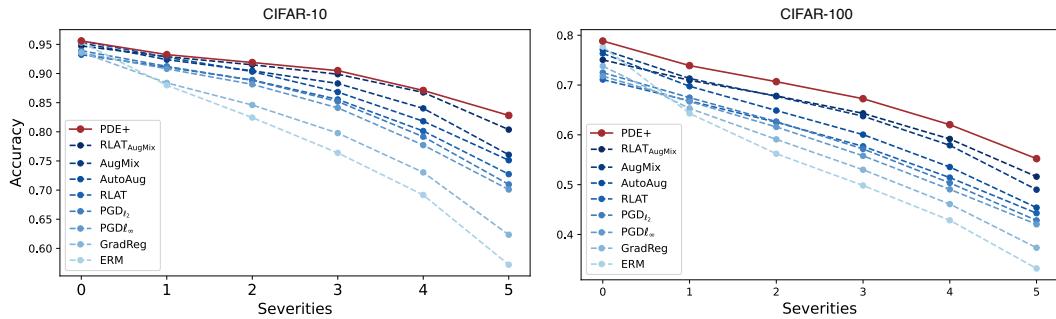


Figure 7: Performance comparison of PDE+ and various baselines on CIFAR-10, CIFAR-10-C, CIFAR-100, and CIFAR-100-C under increasing severity levels. All evaluated models employ the ResNet-18 architecture as the foundation. The x-axis denotes severity levels, with 0 symbolizing the uncorrupted (clean) versions of CIFAR-10 and CIFAR-100 datasets. The y-axis illustrates the accuracy of each model.

E.1 Comprehensive Results for Each Corruption Type

This subsection presents the detailed results obtained by PDE+ and a range of baselines when tested on CIFAR-10-C and CIFAR-100-C. The evaluation metric includes accuracy for each individual corruption (all 15 types of corruptions), along with average accuracy, mCE and rmCE for all corruptions, as shown in Table 3. From these results, we derive two key conclusions: (1) our proposed PDE+ exhibits superior performance across the majority of corruption types, and (2) our PDE+ improves performance in terms of all metrics, including accuracy, mCE and rmCE.

E.2 Exploration of Increasing Severity Impact

This subsection provides analysis of the performance of PDE+ and various baselines on CIFAR10, CIFAR-10-C, CIFAR-100, and CIFAR-100-C datasets with increasing severity levels. The results are visually depicted in Fig. 7. From these results, we infer two main conclusions: (1) Our proposed PDE+ model consistently outperforms other methods across all severity levels, and (2) With an increase in the severity level, PDE+ demonstrates resilience and stability, unlike most baselines, which show a significant decline in performance. This effect is particularly evident in the case of the CIFAR-10 dataset.

E.3 Training Curves

This subsection visualizes the training curves for PDE+ on CIFAR-10 and CIFAR-100, facilitating a comprehensive understanding of model behavior throughout the learning phase.

Diffusion Coverage The evolution of diffusion coverage during the training process is depicted in Fig. 8. A discernible trend emerges from the data: as the model converges, the diffusion coverage contracts and gradually stabilizes at a value exceeding 0. This contraction signifies that the model’s representation of augmented data is progressively aligning more closely with the representation of

Table 3: Comprehensive comparison of PDE+ and various baseline models on CIFAR-10, CIFAR-10-C and CIFAR-100, CIFAR-100-C. All evaluated models employ the ResNet-18 architecture as the foundation. Evaluations are based on Accuracy (%) for each individual corruption, as well as Average Accuracy (Acc %), Mean Corruption Error (mCE %), and Relative Mean Corruption Error (rmCE %) for overall performance. The reported performance of our PDE+ reflects the average across five runs with varying seeds, with a maximum standard deviation under 0.1%. The most notable results are indicated with **boldface** for the top performance, and underline for the second-best results achieved by our PDE+.

Method	Whether			Blur			Noise			Digital			Avg					
	Snow	Fog	Frost	Glass	Defocus	Motion	Zoom	Gaussian	Shot	Impulse	Pixel	Bright	Contrast	JPEG	Elastic	Acc(\uparrow)	Acc(\downarrow)	mCE(\downarrow)
CIFAR-10-C																		
ERM	83.20	89.35	80.05	55.02	82.74	78.48	78.79	48.15	60.63	53.04	75.68	94.04	76.34	78.97	85.05	74.63	100	100
GradReg	82.69	84.50	82.65	59.61	83.38	78.86	80.19	63.80	72.27	64.88	79.97	92.14	70.18	84.81	84.51	77.62	96.29	86.09
AutoAug	87.45	92.56	88.23	80.32	91.15	85.41	89.21	66.91	75.45	81.82	81.70	95.16	94.58	83.76	86.95	85.37	61.74	50.48
AugMix	87.53	91.20	87.86	71.47	92.67	89.32	90.76	79.14	85.03	82.11	84.50	94.18	82.83	89.06	86.25	86.25	60.44	49.01
PGD ℓ_∞	86.68	75.11	86.73	76.33	86.51	81.15	85.13	81.89	85.18	72.23	89.73	91.63	57.60	90.18	86.58	82.17	86.53	77.90
PGD ℓ_2	86.68	77.45	86.85	76.99	87.12	82.53	86.18	81.28	84.90	72.34	89.87	92.27	63.86	90.38	87.46	83.08	81.06	72.07
RLAT	86.13	76.62	87.15	79.94	82.38	86.02	84.69	84.69	87.10	76.42	89.52	91.84	62.50	90.39	87.39	83.67	80.98	66.71
RLAT _{AM}	88.28	88.37	89.18	79.40	92.67	90.03	92.03	85.97	89.22	84.39	90.45	93.66	79.79	90.47	90.32	88.28	55.60	40.20
PDE+	90.09	93.56	89.82	76.32	94.12	91.91	92.81	<u>83.59</u>	<u>87.63</u>	85.10	86.05	<u>94.84</u>	<u>93.22</u>	87.15	90.41	89.11	48.07	33.83
CIFAR-100-C																		
ERM	55.40	64.65	50.49	24.18	60.03	55.45	53.60	23.02	31.44	25.48	52.45	73.72	55.38	52.31	61.50	49.27	100	100
GradReg	55.87	56.40	55.87	28.63	60.15	55.06	55.98	37.12	45.42	35.20	59.40	69.91	45.98	60.44	52.16	52.16	96.95	80.79
AutoAug	60.44	67.63	59.55	46.71	67.04	57.56	62.84	31.86	41.16	58.85	59.30	75.24	74.59	55.60	62.47	58.72	83.12	60.58
AugMix	62.70	67.19	59.73	42.94	76.14	67.79	70.54	48.38	55.85	57.50	62.86	73.60	60.30	61.92	67.42	61.93	77.51	
PGD ℓ_∞	59.61	47.17	59.55	45.86	61.52	55.63	59.55	48.42	53.83	35.37	67.05	67.84	36.37	66.59	61.15	55.03	93.49	68.73
PGD ℓ_2	59.51	48.27	59.28	45.51	61.64	56.02	59.26	51.03	56.52	41.68	67.33	68.51	38.54	66.94	61.41	56.09	91.65	68.58
RLAT	57.81	47.72	57.73	50.73	60.79	55.01	58.33	50.73	59.62	46.97	66.39	66.42	37.65	66.08	60.92	56.54	91.98	64.16
RLAT _{AM}	62.45	60.36	50.49	48.61	70.98	66.94	69.19	56.68	61.76	56.08	69.40	71.55	52.94	61.98	66.70	62.77	77.38	49.66
PDE+	66.68	69.84	63.15	46.99	76.50	73.08	74.38	50.80	58.81	61.48	68.26	76.19	66.57	61.59	69.98	65.62	69.68	47.45

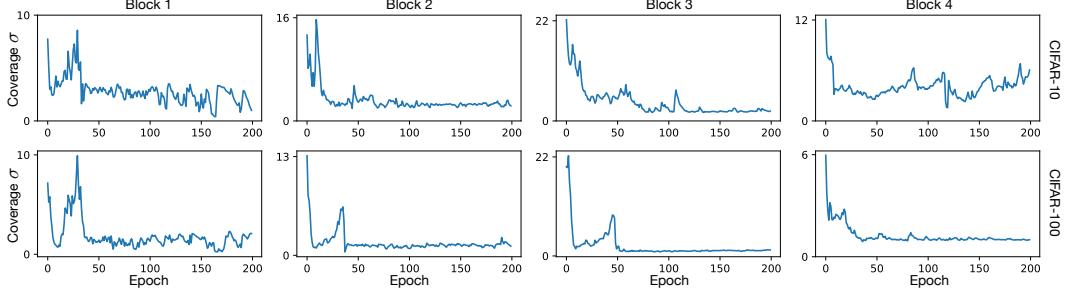


Figure 8: Evolution of diffusion coverage for PDE+ on CIFAR-10 and CIFAR-100 throughout the training phase. The employed backbone is ResNet-18. The x-axis denotes the progression in training epochs, while the y-axis depicts the mean value of diffusion coverage, symbolized as σ .

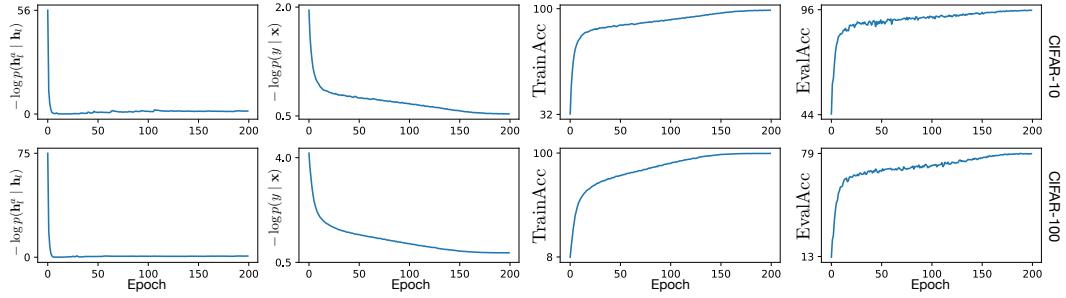


Figure 9: Evolution of loss and accuracy for PDE+ on CIFAR-10 and CIFAR-100 throughout the training phase. The employed backbone is ResNet-18. The x-axis denotes the progression in training epochs, while the y-axis represents different metrics in each column. The first column corresponds to the GaussianNLL loss, the second column to the CrossEntropy loss, the third column is the training accuracy, and the final column is the evaluation accuracy.

the original data, thus indicating an increase in model smoothness across semantically similar areas of entire data space.

Loss and Accuracy The progression of both loss and accuracy throughout the training process is illustrated in Fig. 9. Both the GaussianNLL loss and CrossEntropy loss display a steady decline, simultaneously the training and evaluation accuracy exhibit a continual ascension. This parallel trend affirms the model’s progressive learning efficiency and its growing ability to generalize from the training to the evaluation phase.

E.4 Analysis for Diffusion Coverage of Unseen Distributions

To further validate our generalizability on completely unobserved distribution, we conduct experiments concerning the diffusion coverage of unseen distribution, i.e., probability of samples under training diffusion distribution, where the sample is augmented by training-unseen augmentations. The experiment is based on the 2σ rule of the Gaussian distribution. The implementation proceeds as follows:

During the training phase, we used a single augmentation method op_{train} , while a distinctive technique $op_{test} \neq op_{train}$ was used during testing. Given a model f trained on op_{train} and a sample x . We apply op_{test} to x to derive an augmented sample x' . Inputting x into f gives μ_l and σ_l , the representation and the diffusion range of layer l . Inputting x' into f gives h_l , the representation of layer l . For each layer, we judge whether h_l is a normal sample under the distribution $N(\mu_l, \sigma_l)$ based on the 2σ rule of the Gaussian distribution. The distance-sigma ratio $\frac{|h_l - \mu_l|}{\sigma_l}$ represents the degree of normal sampling. A smaller value indicates a higher probability that h_l is a normal sample under the distribution $N(\mu_l, \sigma_l)$. If the ratio exceeds 2, it suggests that the sample has less than a 5% chance of being normal and is thus likely to be an outlier.

We do not directly calculating the probability $p(h_l; \mu_l, \sigma_l)$ due to the continuous nature of our distribution, which encompasses an infinite sample space. For any specific sample, the probability

approximates zero. Therefore, given the probability density function $d(x)$ and the specific sample h_l , we can only obtain the probability density $d(h_l)$, reflecting the relative density compared to other samples rather than an absolute probability. As such, the most effective method to ascertain whether h_l can be generated by this distribution is based on the 2σ rule.

The complete changing curve of the ratio $\frac{|h_l - \mu_l|}{\sigma_l}$ throughout the training process is provided in Fig. 5. Each row represents a distinct type of training augmentation: brightness, Gaussian, and rotation from top to bottom. Each column is corresponding to each layer of neural network. Each sub-figure includes 3 lines corresponding to the changing trends of ratios during training for test samples generated by 3 different test augmentations. The red dashed line marks the position where ratio is 2

As the training gradually stabilizes, two observations can be noted: (1) Even when training is conducted on a single augmentation completely unseen during the testing phase, the test-phase augmentation samples still have a high likelihood of being normal samples within the model distribution, as they fall within 2σ or even smaller. (2) Different augmentation techniques offer varying capabilities of covering unseen distributions. Of the three augmentations experimented in this study, rotation demonstrated the strongest capability, as it swiftly achieved a lower ratio.

F Settings

This section provides a detailed description of the experimental setups and configurations adopted in this study. These include: (1) Datasets summarized in Appendix F.1; (2) Baseline models against which we benchmarked our proposed model, as is introduced in Appendix F.2; (3) Metrics employed to gauge performance, as is introduced in Appendix F.3; (4) Hyper parameters selected to optimize our model, as is summarized in Appendix F.4. (5) Computing Resources summarized in Appendix F.5.

F.1 Datasets

We perform experiments on 7 common image classification datasets including CIFAR-10, CIFAR-100, Tiny-ImageNet, CIFAR-10-C, CIFAR-100-C, Tiny-ImageNet-C and PACS.

Dataset of Original Distribution Nature distribution of CIFAR-10, CIFAR-100 [40] and Tiny-ImageNet[43] are datasets of original(clean) distribution. CIFAR-10 and CIFAR-100 datasets consist of 60,000 color images, each of size 3x32x32 pixels. CIFAR-10 is categorized into 10 distinct classes with 6000 images per class. CIFAR-100 is more challenging, as these images are distributed across 100 classes, with 600 images per class. Tiny-ImageNet datasets consist of 110,000 color images, each of size 3x64x64 pixels, which are categorized into 200 distinct classes with 550 images per class. Both CIFAR-10 and CIFAR-100 are subdivided into a training set of 50,000 images and a test set of 10,000 images. Tiny-ImageNet is subdivided into a training set of 100,000 images and a test set of 10,000 images.

Dataset of Corrupted Distributions CIFAR-10-C, CIFAR-100-C and Tiny-ImageNet-C[31] are variants of the original CIFAR-10, CIFAR-100 and Tiny-ImageNet datasets that have been artificially corrupted into 19 types of corruptions at 5 levels of severity, resulting in 95 corrupted versions of the original test set images. The corruptions include 15 main corruptions: Gaussian noise, shot noise, impulse noise, defocus blur, glass blur, motion blur, zoom blur, snow, frost, fog, brightness, contrast, elastic, pixelation, and JPEG. Both datasets also contain 4 corruptions that are not commonly used: speckle noise, Gaussian blur, spatter, saturation. All these corruptions are simulations of shifted distributions that models might encounter in real-world situations.

Datset of PACS PACS[45] is an image dataset popular used in domain generalization and transfer learning. It consists of 4 domains, namely Photo (1,670 images), Art Painting (2,048 images), Cartoon (2,344 images) and Sketch (3,929 images). Each domain contains 7 categories.

F.2 Baselines

The various baseline models we selected for comparison with our PDE+ can be categorized into five distinct groups, each representative of a specific approach or methodology:

Table 4: Summary of Original & Corruption Datasets

Dataset	#Train	#Test	#Corr.	#Severity	#Class.
CIFAR-10	50,000	10,000	1	1	10
CIFAR-100	50,000	10,000	1	1	100
Tiny-ImageNet	100,000	10,000	1	1	200
CIFAR-10-C	-	950,000	15(+4)	5	10
CIFAR-100-C	-	950,000	15(+4)	5	100
Tiny-ImageNet-C	-	750,000	15	5	200

Table 5: Summary of PACS Datasets

Domain	#Sample	#Class	Size
Photo	1,670	7	3x227x227
Art	2,048	7	3x227x227
Cartoon	2,344	7	3x227x227
Sketch	3,929	7	3x227x227

Standard Training This category is represented by the foundational ResNet-18 architecture [29]. It serves as a reference point for the rest of the methodologies, showcasing the results of a model trained without any specific regularization, noise injection, or augmentation.

Lipschitz Continuity Methods within this category impose model smoothness directly via the Lipschitz continuity principle. For instance, the work presented in [15] achieves this smoothness by implementing gradient regularization. This approach enhances the stability of the model against small changes in the input space.

Noise Injection This category encompasses methods that utilize various strategies to inject noise. EnResNet [71] improves performance through an ensemble of ResNets with injected noise. RSE [51] presents a framework that injects noise and employs self-ensembling during testing to enhance model robustness. NFM [48] combines noise injection and Manifold Mixup [70] to expand the generalization capacity of the model.

Data Augmentation This group consists of methods via various augmentation strategies to diversify and expand the training dataset. Gaussian noise creates minor variations to increase the robustness. Mixup [75] generates synthetic training examples by creating linear interpolations of random pairs of images and their corresponding labels. DeepAug [30] uses a deep generative model to create synthetic training examples, increasing the diversity of the training set. AutoAug [14] utilizes reinforcement learning to discover the best augmentation policies from a search space of possible augmentations, optimizing the model’s validation accuracy. AugMix [32] employs a combination of multiple augmentation transformations and creates a mixture of augmented images, helping the model to generalize better by exposing it to more varied and complex examples.

Adversarial Training This category uses adversarial examples during training to bolster generalization and robustness. PGD [53] generates adversarial examples using an iterative process, thereby strengthening the model’s robustness against adversarial attacks, which has been recently discovered that by judiciously selecting the perturbation radius, it can enhance non-adversarial generalization on common corruptions [37]. RLAT [37] introduces an efficient relaxation to AT via the distance metric of learned perceptual image patch similarity. This approach combined with data augmentation methods can achieve state-of-the-art performance on common corruptions.

All methods employed for comparison are representative within their respective categories and represent state-of-the-art baselines, providing a comprehensive range of approaches against which the performance of our PDE+ model can be evaluated.

F.3 Metrics

Three metrics are typically employed for reporting the performance of a model on CIFAR-10-C and CIFAR-100-C across different corruption types and severity levels: Average Accuracy, Mean Corruption Error (mCE) [31], and Relative Mean Corruption Error (rmCE) [31]. These metrics

provide a comprehensive evaluation of a model’s generalization in handling diverse corruptions and severities, thereby offering a multi-faceted perspective on model performance.

Average Accuracy Average accuracy is the accuracy averaged over all severity levels and corruptions. Consider there are a total of C corruptions, each with S severities. For a model f , let $\mathcal{E}_{s,c}(f)$ denote the top-1 error rate on the corruption c with severity level s averaged over the whole test set,

$$\text{Accuracy}_f = 1 - \frac{1}{C \cdot S} \sum_{c=1}^C \sum_{s=1}^S \mathcal{E}_{s,c}(f). \quad (39)$$

Mean Corruption Error Mean corruption error (mCE) is a metric used to measure the performance improvement of model f compared to a baseline model f_0 . We use ResNet-18 as the baseline model, instead of AlexNet [41] which is traditionally used in [31]. We define mCE_f as follows,

$$\text{mCE}_f = \frac{1}{C} \sum_{c=1}^C \frac{\sum_{s=1}^S \mathcal{E}_{s,c}(f)}{\sum_{s=1}^S \mathcal{E}_{s,c}(f_0)}. \quad (40)$$

Relative Mean Corruption Error Relative mean corruption error (rmCE) is a variant to mCE, which takes the error rate of models trained on natural data distribution into consideration. Denote the error rate of models f trained on natural data distribution into consideration as $\mathcal{E}_{\text{nat}}(f)$ and the error rate of models f_0 trained on natural data distribution into consideration as $\mathcal{E}_{\text{nat}}(f_0)$. We define rmCE_f as follows,

$$\text{rmCE}_f = \frac{1}{C} \sum_{c=1}^C \frac{\sum_{s=1}^S \mathcal{E}_{s,c}(f) - \mathcal{E}_{\text{nat}}(f)}{\sum_{s=1}^S \mathcal{E}_{s,c}(f_0) - \mathcal{E}_{\text{nat}}(f_0)}. \quad (41)$$

F.4 Hyper-parameters

In this section, we outline the hyperparameters chosen for our experiments, which are based on empirical evaluations. These settings enable the reproducibility of the results presented in our study.

Table 6: Summary of Hyper-parameters

Data	Epochs	B.S.	Classifier			Diffuser	
			lr	Opt.	Sch.	lr	Opt.
CIFAR-10	200	128	0.05	SGD	CosineLR	0.015	Adam
CIFAR-100	200	128	0.05	SGD	CosineLR	0.010	Adam
Tiny-ImageNet	200	128	0.06	SGD	CosineLR	0.005	Adam
PACS	200	128	0.02	SGD	CosineLR	0.005	Adam

F.5 Computing Resources

All our experiments are performed on RedHat server (4.8.5-39) with Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz4 and 4× NVIDIA Tesla V100 SXM2 (32GB)

G Limitations and Future Explorations

Limitations and Future Works This paper presents two limitations that need to be addressed in future work. Firstly, the implementation is based on residual connected network due to the form of TE, which can be further explored to structures without residual connection. Secondly, we acknowledge that the Gaussian prior used for the implementation can be replaced with other priors, such as Laplace prior [39], which is worthy of further exploration. Lastly, experimental on adversarial samples is worth exploring.

Broader Impacts Our method does not raise concerns regarding negative societal impact, as it primarily focuses on enhancing the generalization performance of neural networks. By improving generalization in unknown real-world scenarios, this approach can contribute to the development of more reliable and trustworthy machine learning models.