Attentive Learning Facilitates Generalization of Neural Networks

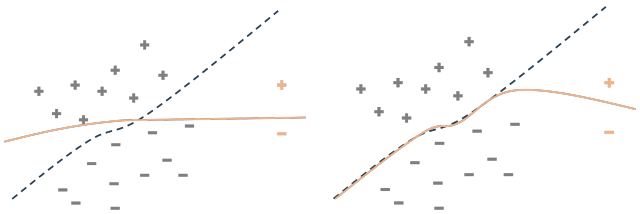
Shiye Lei, Fengxiang He, Haowen Chen, and Dacheng Tao, *Fellow, IEEE*

***Abstract*—This paper studies the generalization of neural networks by examining how a network changes when trained on a training sample with or without out-of-distribution (OoD) examples. If the network's predictions are less influenced by fitting OoD examples, then the network learns attentively from the clean training set. A new notion, *dataset-distraction stability*, is proposed to measure the influence. Extensive CIFAR-10/100 experiments on the different VGG, ResNet, WideResNet, ViT architectures, and optimizers show a negative correlation between the dataset-distraction stability and generalizability. With the distraction stability, we decompose the learning process on the training set $\mathcal{S}$ into multiple learning processes on the subsets of $\mathcal{S}$ drawn from simpler distributions, *i.e.*, distributions of smaller intrinsic dimensions, and further, a tighter generalization bound is derived. Through attentive learning, miraculous generalization in deep learning can be explained and novel algorithms can also be designed. The code is available at https://github.com/LeavesLei/attentive_learning**

***Index Terms*—Explainable AI, Neural Networks, Deep Learning Generalization, Learning Mechanism.**

I. Introduction

NEURAL networks, which are stacked by hundreds of nonlinear mappings, have shown unprecedented prowess in many domain applications of artificial intelligence, such as computer vision [1, 2, 3], natural language processing [4, 5, 6], reinforcement learning [7, 8, 9], self-driving cars [10, 11], and protein prediction [12, 13]. Although many approaches can interpret neural networks' outputs [14, 15], the remarkable success of neural networks is still clouded by the conventional wisdom that large model capacity is typically accompanied by worse performance on unseen data [16], which is known as the famous "Occam's razor" philosophy. To remedy the feeble understanding of the success of deep neural networks, a plethora of works have been done from multiple aspects including model weights, loss landscape, and learning dynamics. For example, Bartlett et al. [17], Neyshabur et al. [18], Chen et al. [19] showed the significant correlation between the spectral norm of weights and generalization performance. Kawaguchi [20], Lu and Kawaguchi [21], Zhou and Liang [22] also proved that all the local minima are equal under some conditions. For training dynamics, Soudry et al. [23], Lyu and Li [24], Ji and Telgarsky [25] revealed the divergence of the weight norm during the training process, which induces the convergence to max-margin solution in networks. In addition, Kalimeris et al. [26] showed that neural networks learn a simple linear function first and then become more nonlinear without forgetting the knowledge of simple functions. Although some progress has been made *w.r.t.* model and algorithm aspects, few works have revolved about the influence of data on the generalizability of networks, while data are the "fuel" to drive model learning and play indispensable roles when understanding the interplay between deep learning models and data.



(a) Non-attentive learning (b) Attentive learning

Fig. 1. An illustration of attentive learning. Initial decision boundary (black curve) is collected by training in-distribution data (+\-). When distraction points (+\-) are injected into the training set, (a) non-attentive learning: distraction points drastically change the in-distribution boundary (orange curve); (b) attentive learning: the in-distribution boundary is less influenced by distraction points.

In this paper, we demonstrate that less interference between in- and out-of-distribution examples contributes to the decent generalizability of neural networks. Specifically, we investigate how interpolating data drawn out of the generated distribution of training data, termed *distraction points*, influences the in-distribution prediction of neural networks. For a distraction set comprising distraction points, if fitting the distraction set hardly affects the predictions on in-distribution data, the learning process is regarded "attentive" to the training data. A new notion, *dataset-distraction stability*, is proposed to characterize attentive learning by the change in the network prediction after learning distraction points.

Extensive experiments are conducted *w.r.t.* different network architectures, optimizers, and training sample sizes on Former/Latter CIFAR-10/100 [1] [27]. The empirical results show a significant negative correlation between the dataset-distraction stability and the generalization performance in neural networks. Moreover, by altering the sample size, label noise ratio, and the

S. Lei and D. Tao are with Sydney AI Centre and School of Computer Science, Faculty of Engineering, The University of Sydney, Darlington NSW 2008, Australia. E-mail: slei5230@uni.sydney.edu.au and dacheng.tao@sydney.ed.u.au.

F. He is with Artificial Intelligence and its Applications Institute, School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, Scotland. E-mail: F.He@ed.ac.uk.

H. Chen is with the Department of Mathematics, ETH Zürich, 8006 Zürich, Switzerland. E-mail: haowchen@student.ethz.ch.

Corresponding author: F. He and D. Tao.

[1]Examples in CIFAR-10/100 belonging to class $0-4/49$ constitute Former CIFAR-10/100, and images in class $5/50-9/99$ form Latter-10/100. Former CIFAR-10/100 and Latter-10/100 are mutual distraction sets.

filter level of the distraction set, we show that although these factors *w.r.t.* distraction sets influence the dataset-distraction stability, the relative magnitude of dataset-distraction stability between different architectures hardly changes. Therefore, the comparison between the distraction stability is not sensitive to the choice of dataset-distraction set, which further supports the relationship between attentive learning and generalization. In addition, we conduct experiments on networks with different layer widths. The result reveals a negative correlation between the dataset-distraction stability and the model capacity, thereby indicating the advantage of overparameterized models in attentive learning.

From a theoretical view, the intrinsic dimension (ID) is employed to suggest the complexity of the data distribution; and we decompose the learning process on the training set $\mathcal{S}$ into multiple learning processes on the subsets of $\mathcal{S}$, which are located on distributions $\mathcal{D}_i$ that are simpler and have smaller ID than $\mathcal{D}$, assisted by the new notion of dataset-distraction stability. This suggests that attentive learning allows the network to learn the decomposed distribution $\mathcal{D}_i$ separately and be less distracted by fitting examples from other decomposed distributions, which alleviates complex prediction rules and thus facilitates the generalization of neural networks. This coincides with the "divide and conquer" philosophy. A tighter upper bound is further proposed based on dataset-distraction stability.

To the best of our knowledge, attentive learning is the first work to demonstrate the generalization of neural networks from the perspective of interference between training data. By establishing correlation between generalizability and the dataset-distraction stability, the surprising generalizability of deep networks could be (partially) attributed to their decent capability of attentive learning. With this, we can further explain many miraculous phenomena in deep learning, such as benign overfitting and out-of-distribution generalization. In addition, by revealing the learning mechanism of networks, attentive learning also has practical applications in designing novel active learning algorithms.

II. Related Works

Generalization in neural networks. Although neural networks are extremely overparameterized and can easily interpolate training data [28, 29, 30], they demonstrate excellent generalizability in many application domains, and the reason has been investigated from multiple perspectives, such as network parameters and learning dynamics. Lu et al. [29, 30] improved the network generalizability by injecting noise in the intermediate feature space as regularization. Wei et al. [31] studied network generalizability in the singular learning theory framework. Keskar et al. [32], He et al. [33] showed that the stochastic gradient descent (SGD) algorithm favors flatter minima in the loss surface, which implies better generalization. In addition, Kawaguchi [20], Lu and Kawaguchi [21], Zhou and Liang [22] proved that there are no spurious local minima for linear networks. Nevertheless, this elegant property does not hold for general networks where nonlinear activation functions are involved [34]. Apart from research on the loss landscape, the implicit bias of neural networks has also been explored in the overparameterized regime. Soudry et al. [23] showed that the overparameterized networks converge to the max-margin solution when the training data are linear-separable. Some follow-up research has also been conducted along this line [24, 25, 35, 36, 37].

There are some works focused on investigating interesting learning properties of networks [38].For example, Belkin et al. [39], Nakkiran et al. [40] revealed the double descent phenomenon in deep learning, *i.e.*, the test loss curve undergoes a second decline with increasing training time when model become overparameterized. In addition, Rahaman et al. [41], Xu et al. [42] discovered that neural networks fit the low-frequency information in the training data first, which constrains the complexity of the learned models. Kalimeris et al. [26] showed that the network gradually learns a more complex function, while the knowledge contained by former functions will not be discarded, during the training process. He et al. [43] showed that significant redundancy existed in the capacity neural network, and the finding may reconcile the mysterious conflict between the extremely large size and excellent generalizability found in deep learning. Wei et al. [44] also studied the influence of out-of-distribution data on generalization by showing that training noisy open-set examples helps alleviate overfitting in inherent label-noise scenarios.

Recently, neural networks have been considered to possess simplicity bias according to the empirical finding that neural networks merely rely on the most discriminative or simplest features to construct the decision boundaries [45, 46]. Following the decision boundary, Lei et al. [47] showed the negative correlation between the variability of the decision boundaries and the generalization when the network was repeatedly trained. Instead, our approach is different from former methods in that it considers the stability of the decision boundary when some distraction points are injected into the training set, and we empirically and theoretically show the correlation between the stability and generalization performance in neural networks.

Intrinsic dimension. Although modern data such as images possess an extremely high extrinsic dimension ($3 \times 32 \times 32 = 3,072$ dimension for the CIFAR-10 dataset), many real-world datasets are conjectured to lie on or around a low-dimensional manifold, *i.e.*, possess a low ID, which is the minimum number of required variables to describe the manifold. This hypothesis has been empirically verified on many natural image datasets [48, 49, 50, 51, 52]. Recently, Pope et al. [53] showed that the ID of datasets has more influence on the generalization performance of neural networks than the original input dimension of datasets, thereby alleviating the curse of dimension. In addition to the global ID, some studies have further considered the variation in the ID within the data manifold or distribution, and they have used local ID to characterize different parts of data distributions [54, 55, 56, 57, 58, 59].

Although estimating the ID is difficult due to the extremely high dimension and sparse sampling, many methods have still been proposed to tackle the problem. Project method searches for the suitable subspace to project the data with less information loss, for instance preserving pairwise distance information between data points [49, 60, 61] or minimizing

the projection error [62]. Some approaches also estimate the ID from the perspective of fractals. Specifically, researchers have counted the number of points within a d-dimensional ball of radius r, and the ID is estimated via the growth rate of this number [63, 64, 65]. In addition, nearest neighbor-based approaches achieve this estimation by leveraging the correlation between the nearest-neighbor statistics and ID [66, 67, 68].

In theory, Narayanan and Mitter [69] proved that the number of required examples to learn a distribution grows exponentially with its intrinsic dimension. Nakada and Imaizumi [70] also derived a generalization bound *w.r.t* ID and does not depend on the original input dimension [71].

III. Preliminaries

$\mathbf{x}$ and y are denoted the observations *w.r.t.* the random variables X and Y, respectively. $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ is the training set, where $\mathbf{x}_i \in [0,1]^n$, n is the dimension of the input data[2], $y_i \in \{1, \ldots, k\}$, k is the number of classes, and $m = |\mathcal{S}|$ is the training sample size. We assume that $(\mathbf{x}_i, y_i)$ are independent and identically distributed (i.i.d.) random variables drawn from the data generating distribution $\mathcal{D}$. The classifier is denoted $f_{\boldsymbol{\theta}}(\mathbf{x}) : [0,1]^n \to \mathbb{R}^k$, which is a neural network parameterized by $\boldsymbol{\theta}$. The output of $f_{\boldsymbol{\theta}}(\mathbf{x})$ is a k-dimensional vector and is assumed to be a discrete probability density function. Let $f_{\boldsymbol{\theta}}^{(i)}(\mathbf{x})$ be the i-th component of $f_{\boldsymbol{\theta}}(\mathbf{x})$, we define $f_{\boldsymbol{\theta}}(\mathbf{x}) = \arg\max_i f_{\boldsymbol{\theta}}^{(i)}(\mathbf{x})$ to denote the prediction made by $f_{\boldsymbol{\theta}}$ on $\mathbf{x}$ for simplicity if there is no ambiguity. For the stochastic learning algorithm $\mathcal{A}$, let $\mathcal{A}(\mathcal{S})$ denote the distribution returned by $\mathcal{A}$ and leveraged on the training set $\mathcal{S}$. Then, we define the random classifier $f_{\mathcal{A}(\mathcal{S})} = \{f_{\boldsymbol{\theta}} | \boldsymbol{\theta} \sim \mathcal{A}(\mathcal{S})\}$. $0-1$ loss is employed in this paper, and the expected risks in terms of $\boldsymbol{\theta}$ and $\mathcal{A}(\mathcal{S})$ are defined as follows:

$$\mathcal{R}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbb{I}\left(y \neq f_{\boldsymbol{\theta}}(\mathbf{x})\right)\right] \tag{1}$$

and

$$\mathcal{R}_{\mathcal{D}}(\mathcal{A}(\mathcal{S})) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\mathbb{E}_{\boldsymbol{\theta}\sim\mathcal{A}(\mathcal{S})}\left[\mathbb{I}\left(y \neq f_{\boldsymbol{\theta}}(\mathbf{x})\right)\right], \tag{2}$$

Here, $\mathbb{I}(\cdot)$ is the indicator function.

A. Distraction Point and Set

The data generating distribution $\mathcal{D} = \mathcal{D}_X \times \mathcal{D}_Y$ can be represented by the joint probability mass function $P(X,Y) = P(X)P(Y|X)$, where $P(X)$ is the data distribution $\mathcal{D}_X$[3]. Then we define the distraction point *w.r.t.* the data generating distribution $\mathcal{D}$ as follows.

Definition 1 (Distraction point)**.** *For the data generating distribution $\mathcal{D} = P(X)P(Y|X)$, $(\mathbf{x}, y)$ is a distraction point of $\mathcal{D}$ if $P(\mathbf{x}) = 0$.*

According to the definition, the distraction point cannot be generated by the distribution $\mathcal{D}$, even with unlimited sampling. After defining the distraction point, we have the following remark:

[2]The input domain of $[0,1]^n$ can be easily achieved via normalization.

[3]The data distribution $\mathcal{D}_X = P(X)$ is also referred to as a covariate distribution in the transfer learning literature.

Remark 1. *(1) Whether the point $(\mathbf{x}, y)$ is a distraction point of $\mathcal{D} = \mathcal{D}_X \times \mathcal{D}_Y$ only depends on $\mathbf{x}$ and the data distribution $\mathcal{D}_X$. Hence, training examples with noisy labels are not distraction points.*
(2) For some distributions, such as the Gaussian distribution in which $P(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathbb{R}^n$, the definition of distraction point can be modified into the threshold-version: $(\mathbf{x}, y)$ is considered a distraction point of $\mathcal{D}$ if $P(\mathbf{x}) < \epsilon$, where ϵ is the threshold.

With the definition of the distraction point, we further introduce the notion of the distraction set.

Definition 2 (Distraction set)**.** *$\mathcal{S}'$ is a distraction set of the distribution $\mathcal{D}$ if $(\mathbf{x}, y)$ is a distraction point of $\mathcal{D}$ for all $(\mathbf{x}, y) \in \mathcal{S}'$.*

We denote $\mathcal{S}' \not\subset \mathcal{D}^{|\mathcal{S}'|}$ if $\mathcal{S}'$ is a distraction set of $\mathcal{D}$ for the sake of simplicity.

B. Intrinsic Dimension

Many works conjecture that the real-world data distribution $\mathcal{D}_X$ can be described by many fewer variables than its extrinsic dimension [49, 72], and the minimal number of variables needed to describe the data distribution is termed as its *intrinsic dimension*. We follow Nakada and Imaizumi [70] to define the intrinsic dimension via the Minkowski dimension, which can describe a broader class of dimension sets and does not require the sets to possess a smooth structure.

Definition 3. *(Minkowski dimension) The Minkowski dimension of a set $E \subset [0,1]^n$ is defined as*

$$dim_M E = \inf\left\{ d^* \geq 0 \mid \limsup_{\varepsilon \downarrow 0} \mathcal{N}(E, \varepsilon)\varepsilon^{d^*} = 0 \right\}, \tag{3}$$

where $\mathcal{N}(\Omega, \epsilon)$ is the fewest number of ϵ-balls that cover Ω in terms of $\|\cdot\|_\infty$ with $\epsilon > 0$.

Remark 2. *Through taking $\mathcal{N}(E, \varepsilon)\varepsilon^{d^*} = C$, where $C \to 0^+$, we can derive $\mathcal{N}(E, \varepsilon) = C(\frac{1}{\epsilon})^{d^*}$. Therefore, $\mathcal{N}(E, \varepsilon)$ can be regarded as the volume of E in terms of the dimension d^* unit length $\frac{1}{\epsilon}$.*

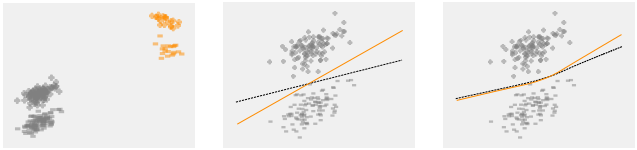
The intrinsic dimension of the data distribution $\mathcal{D}_X$ is defined with the Minkowski dimension:

Definition 4. *(Intrinsic dimension) For the data distribution $\mathcal{D}_X$, the intrinsic dimension of $\mathcal{D}_X$ is defined as*

$$ID(\mathcal{D}_X) = dim_M Supp(\mathcal{D}_X), \tag{4}$$

where $Supp(\mathcal{D}_X)$ is the support of $\mathcal{D}_X$.

Remark 3. *Because the support of $\mathcal{D}_X$ is the closure consisted of all $\mathbf{x}$ for $P(\mathbf{x}) \neq 0$, the uniform noise distribution $P(\mathbf{x}) = \texttt{const} \quad \forall \mathbf{x} \in \mathbb{R}^n$ has the ID of input dimension n. Therefore, for distribution D_X lies on a low dimension manifold, i.e., $ID(\mathcal{D}_X) \leq n$, there exists distraction points $\mathbf{x}$ s.t. $P(\mathbf{x}) = 0$.*



(a) Training points (b) Linear (c) Nonlinear

Fig. 2. (a) In-distribution points (gray) and distraction points (orange) for training MLPs; (b&c) Initial decision boundary (black) and influenced decision boundary (orange) of linear and ReLU MLP.

IV. Attentive Learning

In real-world learning scenarios, training sets are typically contaminated by some distraction points due to either the noisy environment or the lack of data cleaning. These distraction points may interfere with learning from clean training examples and thus affect the in-distribution prediction rule, *i.e.*, the decision boundary. Intuitively, if the impact of distraction points is insubstantial, the model is regarded as learning more attentively from the clean training example. A schematic is shown in Figure 1, where the in-distribution decision boundary is hardly altered with the injected orange distraction points in the training set, thereby showing a desired attentive learning on the in-distribution points (+\-). To better show the attentive learning phenomenon, we conduct a toy example on 2-dimensional points with linear and ReLU multilayer perceptrons (MLPs). The in-distribution points (+\-) and distraction set (+\-) are presented in Figure 2(a), where we plot the initial and influenced in-distraction decision boundaries of linear and nonlinear MLPs in Figures 2(b) and 2(c), respectively. The plots show clear attentive learning, *i.e.*, the in-distribution decision boundary is less influenced by the distraction set in the nonlinear scenario, while this phenomenon does not occur in the linear network.

A. Dataset-distraction Stability

In practice, the influence of injecting a single distraction point is relatively minor and is thus challenging to observe. The distraction set is therefore typically employed to amplify the influence. We formally define the notion of dataset-distraction stability for quantitatively measuring the influence of the distraction set on learning from the source training set.

Definition 5 (Dataset-distraction stability)**.** *Let $f_{\boldsymbol{\theta}}(\mathbf{x}) : [0,1]^n \to \mathbb{R}^k$ be a classifier parameterized by $\boldsymbol{\theta}$. $\mathcal{S} \subset \mathcal{D}^m$ and $\mathcal{S}' \not\subset \mathcal{D}^{m'}$ are the source training set and distraction set of $\mathcal{D}$, respectively. Then, for the stochastic learning algorithm $\mathcal{A}$, the dataset-distraction stability for $f_{\mathcal{A}(\mathcal{S})}$ by the distraction set $\mathcal{S}'$ is*

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\frac{1}{2}\sum_{i=1}^{k}\left|f^{(i)}_{\mathcal{A}(\mathcal{S})}(\mathbf{x}) - f^{(i)}_{\mathcal{A}(\mathcal{S}\cup\mathcal{S}')}(\mathbf{x})\right|\right], \tag{5}$$

where $f^{(i)}_{\mathcal{A}(\mathcal{S})}(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\theta}\sim\mathcal{A}(\mathcal{S})}\left[\mathbb{I}\left(f_{\boldsymbol{\theta}}(\mathbf{x}) = i\right)\right]$, and the dataset-distraction stability is denoted by $\mathrm{DS}(\mathcal{S}, \mathcal{S}'|\mathcal{D})$.

In light of Definition 5, smaller dataset-distraction stability suggests that the prediction of $f_{\mathcal{A}(\mathcal{S})}$ on $\mathcal{D}$ is less influenced by $\mathcal{S}'$, and thus the network f learns from the source training set $\mathcal{S}$ in a more attentive manner via the algorithm $\mathcal{A}$. To further appreciate this definition, we compare it with the well-known algorithmic stability [73] that measures how the training algorithm is sensitive to in-distribution training points.

Definition 6. *(Hypothesis stability [73]) An algorithm $\mathcal{A}$ has hypothesis stability β w.r.t. if*

$$\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^m, z\sim\mathcal{D}}\left[\left|f^{(y)}_{\mathcal{A}(\mathcal{S})}(\mathbf{x}) - f^{(y)}_{\mathcal{A}(\mathcal{S}^{-i})}(\mathbf{x})\right|\right] \leq \beta \tag{6}$$

holds for all $1 \leq i \leq m$, where $z = (\mathbf{x}, y)$ and $\mathcal{S}^{-i} = \{z_1, ..., z_{i-1}, z_{i+1}, ..., z_m\}$ w.r.t. $\mathcal{S} = \{z_1, ..., z_m\}$.

Eq. 6 shows that hypothesis stability solely focuses on the decision boundary *w.r.t.* the ground truth y, while our distraction stability also consider the influence of training data on other inter-class boundary for a more comprehensive measurement. Moreover, our distraction stability perturbs $\mathcal{S}$ with OoD data other than in-distribution points z. In addition, hypothesis stability is not affected by the choice of the training set $\mathcal{S}$ due to the expectation, while our distraction stability considers both the algorithm $\mathcal{A}$ and the training set $\mathcal{S}$ into consideration. Therefore, compared to hypothesis stability, our new notion measures the capability of attentive learning *w.r.t.* the learning process $\{f, \mathcal{A}, \mathcal{S}\}$ and can also handle the distribution shift scenario that exists widely in the real world.

B. Dataset-distraction Stability and Generalization

In this section, we explore the relationship between the capability of attentive learning and the generalizability of neural networks. To measure the dataset-distraction stability, we start by constructing the distraction set.

Distraction set construction. To construct a distraction set that has a disjoint data distribution *w.r.t.* the source set, CIFAR-10 and CIFAR-100 are divided into two datasets according to the categories, respectively. Specifically, the images belonging to classes $0-4$ of CIFAR-10 form *Former CIFAR-10* dataset, whereas the examples coming from classes $5-9$ of CIFAR-10 constitute *Latter CIFAR-10*. Similarly, we evenly partition CIFAR-100 into *Former CIFAR-100* and *Latter CIFAR-100* according to the class without class overlap. As such, the Former CIFAR-10/100 and Latter CIFAR-10/100 can be considered distraction sets for each other.

We conduct experiments with various network architectures and learning algorithms on Former/Latter CIFAR-10/100. Specifically, VGG-16 [74], ResNet-18 [2], WideResNet-28 [75], and Vision Transformer (ViT) [76], are optimized by vanilla SGD, SGD with momentum, RMSprop [77], and Adam [78]. According to the construction of distraction sets, when Former CIFAR-10/100 is the source training set $\mathcal{S}$, Latter CIFAR-10/100 is the corresponding distraction set $\mathcal{S}'$, and vice versa. For each setting of (architecture, optimizer, dataset), the network f is repeatedly trained 10 times with the same initialization to estimate the random classifier $f_{\mathcal{A}(\mathcal{S})}$.

After training networks on the source training set $\mathcal{S}$ and the distracted training set $\mathcal{S} \cup \mathcal{S}'$, the dataset-distraction stability $\mathrm{DS}(\mathcal{S}, \mathcal{S}'|\mathcal{D})$ can be estimated on the corresponding source test set in terms of $\mathcal{S}$. Then, by calculating the test error $\mathcal{R}_{\mathcal{D}}(\mathcal{A}(\mathcal{S}))$ of $f_{\mathcal{A}(\mathcal{S})}$ on the source test set, Figure 3 shows the scatter plots

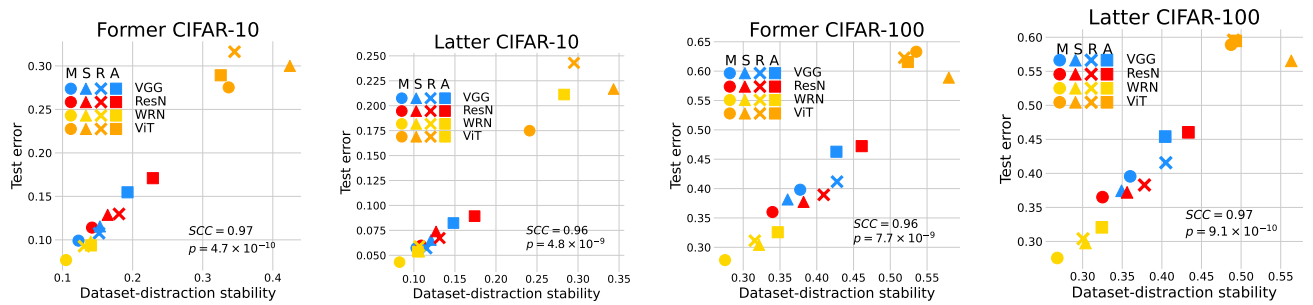
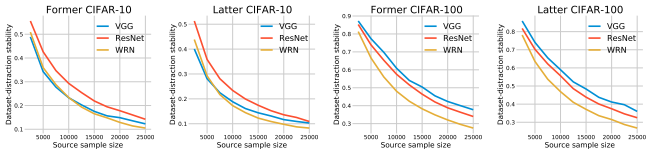
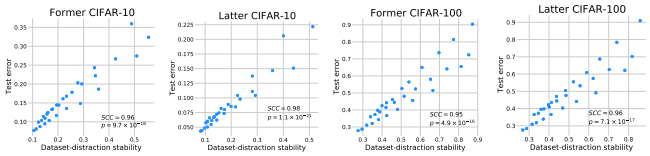


Fig. 3. Scatter plots of dataset-distraction stability to test error with different architectures and optimizers on Former/Latter CIFAR-10/100. The colors of blue, red, yellow, and orange points denote the architectures of VGG-16 (VGG), ResNet-18 (ResN), WideResNet-28 (WRN), ViT, respectively. The shapes of ●, ▲, ✖, and ■ designate the learning algorithms of SGD with momentum (M), vanilla SGD (S), RMSprop (R), and Adam (A), respectively.

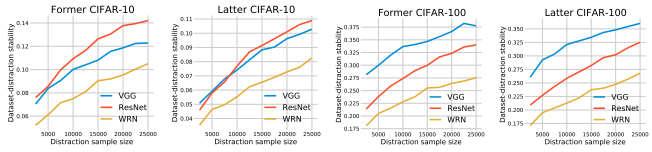


(a) Dataset-distraction stability vs. source sample size

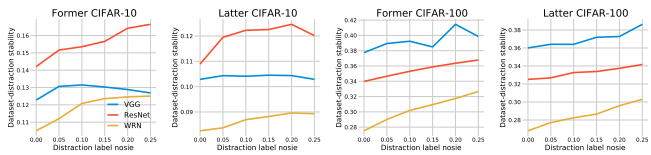


(b) Test error vs. dataset-distraction stability

Fig. 4. The relationship between dataset-distraction stability and source sample size on Former/Latter CIFAR-10/100: (a) Plots of dataset-distraction stability as a function of source sample size; (b) Scatter plots of test error to dataset-distraction stability. The points are collected from various source sample sizes and network architectures.



(a) Distraction sample size vs. dataset-distraction stability



(b) Distraction label noise vs. dataset-distraction stability

Fig. 5. Plots of dataset-distraction stability that acts as as a function of (a) distraction sample size and (b) distraction label noise on Former/Latter CIFAR-10/100.

of (dataset-distraction stability, test error). For each plot, we also calculate Spearman's rank-order correlation coefficients (SCCs) and the corresponding p value of the collected data to investigate the statistical significance of the correlations; please see the bottom right corner of the plots. From the plots, several observations can be derived: (1) SGD with momentum possesses a better generalization performance and smaller dataset-distraction stability than other learning algorithms in most cases; (2) WideResNet has both lower test error and dataset-distraction stability than VGG, ResNet, and ViT; (3) ViT has both the highest test error and dataset-distraction stability; and (3) there is a positive correlation between test error and dataset-distraction stability ($SCCs > 0.9$), and the correlation is statistically significant ($p < 0.005$)[4]. Therefore, we rationally propose the following conjecture:

Hypothesis 1. *With fixed $\mathcal{D}$ and $\mathcal{S}' \not\subset \mathcal{D}^{|\mathcal{S}'|}$, the network f trained on $\mathcal{S} \subset \mathcal{D}^m$ with smaller dataset-distraction stability possesses better generalization performance.*

[4]The definition of "statistically significant" has various versions, such as $p < 0.05$ and $p < 0.1$. This paper uses a more rigorous definition ($p < 0.005$).

C. Distraction stability and Source Sample Size

In this section, we investigate the influence of the source training sample size $|\mathcal{S}|$, which is an important factor affecting the generalization performance, on the dataset-distraction stability. To this end, we change the source training sample size $|\mathcal{S}|$, and the distraction set $\mathcal{S}'$ is kept invariant. Then, the plots of dataset-distraction stability *w.r.t.* different source sample sizes can be calculated, as shown in Figure 4(a). The plots suggest that dataset-distraction stability decreases with increasing source sample size. In other words, a larger sample size helps the network learn more attentively about the distribution $\mathcal{D}$. Moreover, we collect the points of (dataset-distraction stability, test error) from various source sample sizes and architectures, and the SCCs and the corresponding p value are calculated to investigate the statistical significance of the correlations, as shown in Figure 4(b). The scatter plots present a clear positive correlation between test error and dataset-distraction stability ($SCCs > 0.95$), and the correlation is statistically significant ($p < 0.005$), thereby supporting Hypothesis 1.

D. Dataset-distraction Stability and Distraction Set

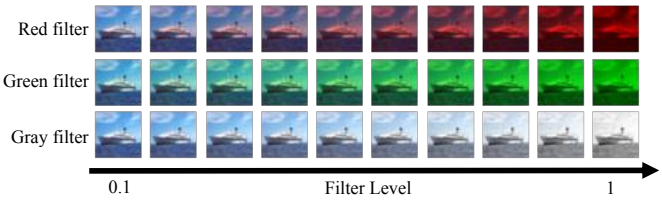
According to Definition 5, dataset-distraction stability is directly influenced by choices of the distraction set $\mathcal{S}'$, including its sample size, label noise ratio, and similarity to

the source training set $\mathcal{S}$. In this section, we aim to answer the following two questions: (1) how does the choice of distraction set influence the dataset-distraction stability and (2) does the positive correlation between dataset-distraction stability and test error still hold regardless of the choice of distraction set? The experiments are performed on four datasets of Former/Latter CIFAR-10/100 with the three VGG, ResNet, and WideResNet architectures, and the optimizer is SGD with momentum.

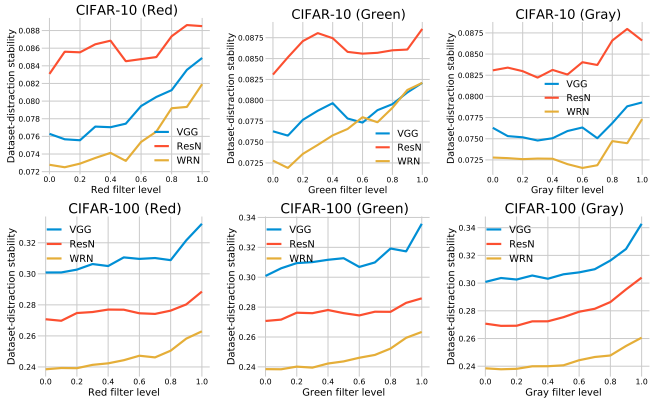
Distraction sample size. By varying the sample size of the distraction set $|\mathcal{S}'|$ and unchanging the source training set $\mathcal{S}$, we plot the change in dataset-distraction stability *w.r.t.* different distraction sample sizes, as shown in Figure 5(a). From the plots, we observe that (1) dataset-distraction stability increases as the distraction sample size increases, which implies that a larger distraction set $\mathcal{S}'$ has a stronger influence on the prediction of $f_{\mathcal{A}(\mathcal{S})}$, and (2) there are few crosses between the curves of different network architectures in these plots, especially when the distraction sample is large. In other words, the distraction sample size does not influence the relative magnitude between the dataset-distraction stability of different network architectures. Therefore, although the distraction sample size directly affects the dataset-distraction stability, a network with better generalization performance always possesses smaller dataset-distraction stability regardless of the sample size of the distraction set. This observation indicates that dataset-distraction stability is a reliable indicator of generalization *w.r.t.* different architectures when the distraction set is fixed.

Label noise alters the conditional label distribution $P(Y|X)$, and noisy data typically require more training time or model capacity for interpolation. We explore how the label noise of the distraction set affects the dataset-distraction stability. Noisy distraction sets with five different label noise ratios are first constructed based on the training set of Former/Latter CIFAR-10/100; the generation of label noise and the complete list of noise ratios can be found in Appendix VII-C4. Then we measure the dataset-distraction stability with different label noise ratios in the distraction set, while the source training set $\mathcal{S}$ is clean and invariant. The results are presented in Figure 5(b). From the plots, we observe that (1) the dataset-distraction stability of VGG plateaus along with the increase in distraction label noise rate in Former CIFAR-10 and Latter CIFAR-10; (2) in most cases, the dataset-distraction stability has a slight increase when the distraction label noise rate increases. In sum, the variation caused by the noisy distraction set on dataset-distraction stability is insubstantial compared to changing the architecture, optimizer, or sample size. In addition, introducing label noise hardly changes the relative magnitude between the dataset-distraction stability of different architectures, which consequently supports our argument about the correlation between attentive learning and generalization in neural networks.

Similarity between $\mathcal{S}$ and $\mathcal{S}'$. We also explore the influence of the similarity between the source training set $\mathcal{S}$ and the distraction set $\mathcal{S}'$ on the dataset-distraction stability. To achieve this investigation, we randomly select $15,000$ images from the training set of CIFAR-10/100 for the construction of distraction sets $\mathcal{S}'$. The remaining unselected training images constitute the source training set $\mathcal{S}$. Then, three different colors (red, green, gray) and 10 different levels of filters are added to the selected images to construct the distraction sets with different similarities. Images in the same distraction set are added with the same filter, *i.e.*, the same color and level; more details can be found in Appendix VII-C4. Figure 6(a) presents the distraction images with different filters. Therefore, a total of 3 colors $\times$ 10 filter_levels $=$ 30 distraction sets are constructed for CIFAR-10 and CIFAR-100.



(a) Images with different filters

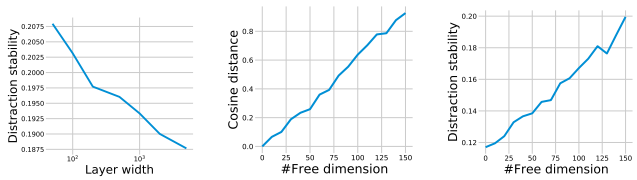


(b) Dataset-distraction stability vs. filter level

Fig. 6. (a) Examples of image with different filters (red, green, and gray); (b) plots of dataset-distraction stability as a function of filter level on CIFAR-10 and CIFAR-100 with different types of filters.

With the constructed source training sets and distraction sets, we can explore the dataset-distraction stability trend in terms of distraction sets with different similarities. The experimental results are presented in Figure 6(b). From the plots, several observations can be obtained: (1) dataset-distraction stability increases with increasing filter level, which indicates that decreased similarity between the source training set and distraction set can improve the dataset-distraction stability, and (2) the curves of different network architectures show a similar trend *w.r.t.* the filter level and have few crosses between them. As such, the similarity between the distraction set and source set also hardly influences the relative magnitude between dataset-distraction stability of different network architectures.

By exploring the influence of distraction sets with different sample sizes, label noise ratios, and filter levels (similarities) on dataset-distraction stability, we show that the choice of distraction set hardly changes the relative magnitude of dataset-distraction stability *w.r.t.* different learning processes when the source training set is fixed. In other words, justifying generalization performance via dataset-distraction stability is not sensitive to the choice of distraction set, which further enhances the connection between generalization and attentive learning in neural networks.



(a) (b) (c)

Fig. 7. Experimental results on high-dimensional data. (a) Plot of dataset-distraction stability as a function of layer width on high-dimensional data; (b) plots of the cosine distance between $\mathbf{u}$ and $\mathbf{u}'$ as a function of the number of free dimensions; and (c) plots of dataset-distraction stability as a function of the number of free dimensions.

E. Other Factors Influencing Attentive Learning

Apart from $\mathcal{S}$ and $\mathcal{S}'$, we also investigate two other essential factors that affect attentive learning in this section: (1) model capacity and (2) the similarity between the source and distraction set. Compared to the similarity investigation in Section IV-D, we more systematically manipulate the similarity between $\mathcal{S}$ and $\mathcal{S}'$ based on 150-dimensional points generated from a Gaussian distribution. The data generation process is presented as follows, and we employ the three-layer ReLU MLP to fit the generated examples. More details can be found in Appendix VII-C5.

High-dimensional data generation. To generate the source training $\mathcal{S}$ and its corresponding test set $\mathcal{S}_{\text{test}}$, we first fix a normal vector $\mathbf{u} \sim [-\frac{1}{6}, \frac{1}{6}]^{150}$ whose elements are independently sampled from the uniform distribution $\mathbf{U}(-\frac{1}{6}, \frac{1}{6})$. To obtain the in-distribution point $(\mathbf{x}, y)$, we sample $\mathbf{z} \sim \mathbf{N}(0, \mathbf{I})$ and $y \in \{-1, 1\}$, and then $\mathbf{x} = \mathbf{z} + y\mathbf{u}$. The generation of $\mathcal{S}'$ is the same as that of $\mathcal{S}$ except another normal vector $\mathbf{u}'$ sampled from the uniform distribution is used. Due to the high dimension of data, the randomly generated $\mathbf{u}$ and $\mathbf{u}'$ are commonly nearly orthogonal. Thus the decision rules of $\mathcal{S}$ and $\mathcal{S}'$ are very dissimilar.

Effect of layer width. We train three-layer MLPs with seven different widths on the generated $\mathcal{S}$ and $\mathcal{S} \cup \mathcal{S}'$. The dataset-distraction stability is then measured on $\mathcal{S}_{\text{test}}$. As shown in Figure 7(a), the distraction stability decreases as the layer width increases, which suggests that the layer model capacity helps sidestep the interference between fitting examples drawn from different distributions and contributes to better attentive learning.

Effect of the distraction set. To manipulate the similarity between $\mathcal{S}$ and $\mathcal{S}'$, we can control the cosine distance between their normal vectors $\mathbf{u}$ and $\mathbf{u}'$. In practice, we specify some dimensions of $\mathbf{u}'$ as the *free dimension*: elements in the free dimension of $\mathbf{u}'$ are randomly sampled from $\mathbf{U}(-\frac{1}{6}, \frac{1}{6})$, while elements in other dimensions are the same as those in $\mathbf{u}$. Hence, fewer free dimensions of $\mathbf{u}'$ imply a smaller cosine distance between $\mathcal{S}$ and $\mathcal{S}'$, as shown in Figure 7(b). By changing the number of free dimensions in $\mathbf{u}'$, we can generate multiple distraction sets with different similarities compared to the source set. Then, we train MLPs with a width of 200 on these distraction and source sets and calculate the dataset-distraction stability on $\mathcal{S}_{\text{test}}$, and the result is presented in Figure 7(c). We observe that more free dimensions of $\mathbf{u}'$ contribute to increased dataset-distraction stability, which suggests that a more different distraction set has a larger impact on learning from the source set; and the result is consistent with the observations in Section IV-D.

In sum, we recall our observations in this section as follows:

- Dataset-distraction stability is significantly correlated with generalization performance of neural networks by varying network architectures and optimizers (Section IV-B);
- Larger sample size help decrease dataset-distraction stability and also test error (Section IV-C);
- The relative magnitude of dataset-distraction stability is hardly changed by varying the distraction set (Section IV-D);
- The dataset-distraction stability increase along with the less similarity between the source set and distraction set (Sections IV-D and IV-E);
- Wide neural networks help decrease dataset-distraction stability (IV-E).

V. Theoretical Evidence

In this section, we explore and develop the theoretical functions for our dataset-distraction stability. Our theoretical analysis can be divided into two steps: we first provide the theoretical guarantee *w.r.t* the complexity of data distribution, which is measured by intrinsic dimension. Then, we employ the notion of dataset-distraction stability to decompose the learning process *w.r.t* $\mathcal{D}$ into multiple learning process *w.r.t.* multiple simpler sub-distribution of $\mathcal{D}$. In the end, we can derive a generalization bound *w.r.t.* dataset-distraction stability and intrinsic dimensions of these decomposed data distributions.

Then, an upper bound *w.r.t.* the intrinsic dimension of $\mathcal{D}_X$ can be derived as follows.

Lemma 1. *Let $f : [0,1]^n \to \mathbb{R}^2$ be a binary classifier. For the data generating distribution $\mathcal{D} = \mathcal{D}_X \times \mathcal{D}_Y$, $\mathcal{S} \in \mathcal{D}^m$ is the training set with size m. If the training error $\mathcal{R}_{\mathcal{S}}(\mathcal{A}(\mathcal{S})) = 0$ and intrinsic dimension ID$(\mathcal{D}_X) = d$, then for the random classifier $f_{\mathcal{A}(\mathcal{S})}$, with probability of at least $1 - 2\exp\left(-m^{d/(2\gamma+d)}\right)$ we have*

$$\mathcal{R}_{\mathcal{D}}\left(\mathcal{A}(\mathcal{S})\right) \leq C m^{-2\gamma/(2\gamma+d)} \left(1 + \log m\right)^2 \tag{7}$$

for any $m \geq N$ with a sufficiently large N, where C is a constant and γ is a degree of smoothness that only depends on $\mathcal{D}$.

This lemma shows that the expected risk is upper bounded by the intrincsic dimension. The proof is for the binary classification problem and inspired by Nakada and Imaizumi [70], and the detail is given in Appendix VIII-A.

Through the definition of dataset-distraction stability, we can derive the bound for the difference between the expected risk of $f_{\mathcal{A}(\mathcal{S} \cup \mathcal{S}')}$ and $f_{\mathcal{A}(\mathcal{S})}$ on the distribution $\mathcal{D}$.

Proposition 1. *Let $f : [0,1]^n \to \mathbb{R}^2$ be a binary classifier. $\mathcal{S} \in \mathcal{D}^m$ is the source training set. If the dataset-distraction stability for $f_{\mathcal{A}(\mathcal{S})}$ by the distraction set $\mathcal{S}'$ is β, then*

$$\left|\mathcal{R}_{\mathcal{D}}\left(\mathcal{A}\left(\mathcal{S} \cup \mathcal{S}'\right)\right) - \mathcal{R}_{\mathcal{D}}\left(\mathcal{A}\left(\mathcal{S}\right)\right)\right| \leq \beta. \tag{8}$$

TABLE I
ESTIMATED INTRINSIC DIMENSIONS FOR CIAFR-10/100 AND SUBSETS OF CIFAR-10/100 (SUB-CIFAR-10/100) USING DIFFERENT k'S. WE PRESENT BOTH AVERAGED ESTIMATED ID AND THE LARGEST ID (IN BRACKET) *w.r.t.* SUBSETS.

Dataset	Sub-CIFAR-10	CIFAR-10	Sub-CIFAR-100	CIFAR-100
MLE ($k=3$)	10.63 (11.71)	13.30	9.40 (10.88)	12.18
MLE ($k=4$)	14.70 (16.16)	17.56	12.93 (14.81)	15.99
MLE ($k=5$)	16.81 (18.66)	19.28	14.77 (16.89)	17.57

The proof is given in Appendix VIII-B. According to Proposition 1, low dataset-distraction stability implies that the distraction set $\mathcal{S}'$ will have less impact on the expected risk $\mathcal{R}_{\mathcal{D}}\left(\mathcal{A}\left(\mathcal{S}\right)\right)$ when $\mathcal{S}'$ is injected into the training set. Then we show the correlation between dataset-distraction stability and generalization via the notion of intrinsic dimension. Because many real-world datasets contain regions with different IDs [56], we further propose the following conjecture:

Hypothesis 2. *(Data distribution decomposition) The data distribution $\mathcal{D}_X$ can be decomposed by $\mathcal{D}_X = \alpha_1\mathcal{D}_1 + \cdots + \alpha_T\mathcal{D}_T$ such that*
(1) $\sum_i^T \alpha_i = 1$;
(2) $P_i(X)P_j(X) = 0$ for all $1 \le i < j \le T$;
(3) ID$(\mathcal{D}_i) <$ ID$(\mathcal{D})$ for all $1 \le i \le T$.

Hypothesis 2 states that distribution $\mathcal{D}_X$ can be represented by multiple disjoint distributions $\{\mathcal{D}_1, \cdots, \mathcal{D}_T\}$ with lower intrinsic dimensions, and we provide empirical evidence for this hypothesis. Specifically, we evenly divide CIFAR-10 and CIFAR-100 into 10 subsets according to categories; and then, the maximum likelihood estimation (MLE) of Levina and Bickel [66] is employed to estimate the ID of CIFAR-10/100 and their subsets. Notably, the MLE estimator requires computing the Euclidean distance to the k-th nearest neighbor, and we select multiple values of hyperparameter $k = \{3, 4, 5\}$ for comprehensive ID comparison. As shown in Table I, we observe that the ID of CIFAR-10/100 is always larger than that of their subsets, and Hypothesis 2 is consequently empirically verified.

Through data distribution decomposition, the data generating distribution $\mathcal{D}$ can be decomposed by $\sum_{i=1}^T \alpha_i\mathcal{D}_i$, and the training set $\mathcal{S}$ can also be similarly represented by $\mathcal{S} = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_T$, where $\mathcal{S}_i \subset \mathcal{D}_i^{|\mathcal{S}_i|}$ for $1 \le i \le T$. Hence, the following equation can be derived *w.r.t.* the expected risk of $f_{\mathcal{A}(\mathcal{S})}$:

$$\begin{aligned}\mathcal{R}_{\mathcal{D}}(\mathcal{A}(\mathcal{S})) =& \alpha_1\mathcal{R}_{\mathcal{D}_1}(\mathcal{A}(\mathcal{S}_1 \cup \cdots \cup \mathcal{S}_T)) \\ &+ \cdots + \alpha_T\mathcal{R}_{\mathcal{D}_t}(\mathcal{A}(\mathcal{S}_1 \cup \cdots \cup \mathcal{S}_T)). \end{aligned} \tag{9}$$

According to Proposition 1, we have

$$\mathcal{R}_{\mathcal{D}_i}(\mathcal{A}(\mathcal{S}_1 \cup \cdots \cup \mathcal{S}_T)) \le \mathcal{R}_{\mathcal{D}_i}(\mathcal{A}(\mathcal{S}_i)) + \beta_i \tag{10}$$

for all $1 \le i \le T$, where $\beta_i = \mathrm{DS}(\mathcal{S}_i, \mathcal{S}\backslash\mathcal{S}_i|\mathcal{D}_i)$ is the dataset-distraction stability. Then, the expected risk $\mathcal{R}_{\mathcal{D}}(\mathcal{A}(\mathcal{S}))$ can be bounded by several weighted expected risks $\mathcal{R}_{\mathcal{D}_i}(\mathcal{A}(\mathcal{S}_i))$ *w.r.t.* $\mathcal{D}_i$ with the lower-dimensional data structure plus dataset-distraction stability as follows:

$$\begin{aligned}\mathcal{R}_{\mathcal{D}}(\mathcal{A}(\mathcal{S})) \le& \alpha_1\mathcal{R}_{\mathcal{D}_1}(\mathcal{A}(\mathcal{S}_1)) + \cdots + \alpha_T\mathcal{R}_{\mathcal{D}_T}(\mathcal{A}(\mathcal{S}_T)) \\ &+ \alpha_1\beta_1 + \cdots + \alpha_T\beta_T. \end{aligned} \tag{11}$$

As such, with more attentive learning (smaller β), the network learns the simple decomposed distribution $\mathcal{D}_i$ more separately and is less distracted by training samples in other decomposed distributions, which consequently facilitates the generalization of neural networks. This coincides with the "divide and conquer" philosophy.

Then, with Lemma 1 *w.r.t.* the intrinsic dimension bound, we can derive the following theorem *w.r.t.* the dataset-distraction stability.

Theorem 2. *(Dataset-distraction stability-based bound) Let $f : [0,1]^n \to \mathbb{R}^2$ be a binary classifier. For the data generating distribution $\mathcal{D} = \mathcal{D}_X \times \mathcal{D}_Y$, $\mathcal{S} \in \mathcal{D}^m$ is the source training set. If the training error $\mathcal{R}_{\mathcal{S}_i}(\mathcal{A}(\mathcal{S}_i))$ is zero and $\max_i \mathrm{ID}(\mathcal{D}_i) = d_{\max} < \mathrm{ID}(\mathcal{D}_X) = d$, then with probability of at least $1 - 2\exp\left(-m^{d/(2\gamma+d)}\right)$ we have*

$$\mathcal{R}_{\mathcal{D}}\left(\mathcal{A}(\mathcal{S})\right) \le Cm^{-2\gamma/(2\gamma+d_{\max})}\left(1+\log m\right)^2 + \sum_{i=1}^T \alpha_i\beta_i \tag{12}$$

for any $m \ge N$ with a sufficiently large N, where C is a constant and γ is a degree of smoothness that only depends on $\mathcal{D}$.

The proof is given in Appendix VIII-C. Theorem 2 shows that two parts bound the expected risk. (1) The first term depends on the intrinsic dimension of the decomposed data distribution $\mathcal{D}_i$, which can also be understood as the complexity of the subtasks. By finding a data distribution decomposition with smaller $d_{\max}$, we can get a tighter bound than Lemma 1. (2) The second term contains distraction stability and can be decreased by choosing models and learning algorithms with better capability of attentive learning.

In a more complex scenario in which some decomposed distributions $\mathcal{D}_i$ are still complex (large d_i), but it has a small proportion *w.r.t* $\mathcal{D}$ (small α_i), we can derive the following corollary that replaces $d_{\max}$ with a smaller intrinsic dimension of decomposed distributions:

Corollary 1. *Let $f : [0,1]^n \to \mathbb{R}^2$ be a binary classifier. For the data generating distribution $\mathcal{D} = \mathcal{D}_X \times \mathcal{D}_Y$, $\mathcal{S} \in \mathcal{D}^m$ is the source training set. If the training error $\mathcal{R}_{\mathcal{S}_i}(\mathcal{A}(\mathcal{S}_i))$ is zero and $d_{\max} = d_1 \ge d_2 \ge \cdots \ge d_T$, then with probability of at least $1 - 2\exp\left(-m^{d/(2\gamma+d)}\right)$ we have*

$$\mathcal{R}_{\mathcal{D}}\left(\mathcal{A}(\mathcal{S})\right) \tag{13}$$

$$\le \min_t \left\{ Cm^{-2\gamma/(2\gamma+d_t)}\left(1+\log m\right)^2 + \sum_{i=1}^t \alpha_i + \sum_{i=t}^T \alpha_i\beta_i \right\} \tag{14}$$

for any $m \ge N$ with a sufficiently large N, where C is a constant and γ is a degree of smoothness that only depends on $\mathcal{D}$.

Remark 4. *Corollary 1 is derived from Theorem 2 by maximizing the expected risk on complex decomposed distribution $\mathcal{D}_i$ to its proportion α_i for dropping its large intrinsic dimension d_i.*

VI. Discussion

With attentive learning, we can provide insightful explanations about many miraculous phenomena in deep learning. Moreover, novel algorithms can be designed in many application domains via attentive learning.

Benign overfitting. According to statistical wisdom, overfitting the training points will hurt the generalization due to the increase in the complexity of the prediction rules. However, neural networks often perfectly interpolate the training data while preserving decent generalization performance, which means that overfitting in deep learning is benign and hardly spoils generalization. For the data generating distribution $\mathcal{D}$ that can be decomposed by multiple disjoint distributions of $\{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_T\}$, each decomposed distribution $\mathcal{D}_i$ possesses a lower-dimensional structure of $\mathcal{D}_X^i$ and has simpler prediction rules than $\mathcal{D}$. Then, the predictor can be regarded to interpolate the data drawn from these decomposed distributions. According to the notion of attentive learning, overfitting the points drawn from $\mathcal{D}_i$ has little influence on improving the complexity of the prediction rule such as the decision boundary *w.r.t.* $\mathcal{D}_j$ when $i \neq j$. Therefore, attentive learning can reduce the complexity of prediction rules and thus alleviate detriments caused by the overfitting of networks.

Associated out-of-distribution generalization. Miller et al. [79] empirically showed the linear correlation between out-of-distribution and in-distribution generalization. In the OoD setting, there is a data distribution shift for $P_{\text{OoD}}(X)$ from the source $P(X)$, while the conditional label distribution is typically unchanged, *i.e.*, $P_{\text{OoD}}(Y|X) = P(Y|X)$. We assume that the intersection between $\mathcal{D}_X$ and $\mathcal{D}_X^{\text{OoD}}$ is not empty. Then, the source training set $\mathcal{S} \subset \mathcal{D}^m$ can be partitioned into $\mathcal{S}_{\text{inter}} \cup \mathcal{S}_{\text{intra}}$, where $P_{\text{OoD}}(\mathbf{x}) > 0$ for $\mathbf{x} \in \mathcal{S}_{\text{inter}}$ and $P_{\text{OoD}}(\mathbf{x}) = 0$ for $\mathbf{x} \in \mathcal{S}_{\text{intra}}$. Through the dataset-distraction stability, the expected OoD risk can be represented as follows:

$$\mathcal{R}_{\mathcal{D}_{\text{ood}}}(\mathcal{A}(\mathcal{S})) \leq \mathcal{R}_{\mathcal{D}_{\text{ood}}}(\mathcal{A}(\mathcal{S}_{\text{inter}})) + \text{DS}(\mathcal{S}_{\text{inter}}, \mathcal{S}_{\text{intra}}|\mathcal{D}_{\text{ood}}). \tag{15}$$

The first term of $\mathcal{R}_{\mathcal{D}_{\text{ood}}}(\mathcal{A}(\mathcal{S}_{\text{inter}}))$ can be decreased by learning from the "helpful set" $\mathcal{S}_{\text{inter}}$ because $\mathcal{S}_{\text{inter}}$ is located in $\mathcal{D}_{\text{ood}}$, and attentive learning helps decrease the second term of $\text{DS}(\mathcal{S}_{\text{inter}}, \mathcal{S}_{\text{intra}}|\mathcal{D}_{\text{ood}})$, *i.e.*, it inhibits the impact brought by fitting the "harmful set" $\mathcal{S}_{\text{intra}}$. Considering the finding that attentive learning facilitates in-distribution generalization, the correlation between OoD and in-distribution generalization can therefore be connected via attentive learning.

Active learning. Our study in active learning can be extended to many application domains such as active learning [80], which aims to select examples to label from the unlabeled data pool to save building a training set and efficiently improve generalization performance. Because the calculation of dataset-distraction stability does not require the label, we can label the example $\mathbf{x}$ with a large $\text{DS}(\mathcal{S}, \mathcal{S}'|\mathbf{x})$. Therefore, the dataset-distraction stability can be efficiently decreased by putting a new labeled example $z = (\mathbf{x}, y)$ into the training set. As such, the network learns more attentively from the new training set $\mathcal{S} \cup \{z\}$, and the generalization performance can be accordingly improved.

VII. Additional Experiments Details

A. Datasets

Our experiments are conducted on two public datasets (CIFAR-10 and CIFAR-100 [27]). **CIFAR-10** consists of $50,000$ training images and $10,000$ test images from 10 different classes, and **CIFAR-100** consists of $50,000$ training images and $10,000$ test images from 100 different classes. One can download CIFAR-10 and CIFAR-100 from https://www.cs.toronto.edu/~kriz/cifar.html.

B. Model architectures

We use different neural network architectures in our experiments, including VGG-16, ResNet-18, WideResNet-28, and ViT. The architectures of CNNs are presented in Table II, where fc denotes the fully-connected layer. The ViT architecture has 4×4 patch resolution, 6 layers, 8 heads, hidden dimension of 512, and the MLP dimension of 512.

TABLE II
Detailed model architectures for Former/Latter CIFAR-10/100

VGG-16	ResNet-18	WideResNet-28-10
$(3 \times 3, 32) \times 2$	3×3, 64	3×3, 16
maxpool, 2×2		
$(3 \times 3, 128) \times 2$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 160 \\ 3 \times 3, 160 \end{bmatrix} \times 4$
maxpool, 2×2		
$(3 \times 3, 256) \times 3$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 320 \\ 3 \times 3, 320 \end{bmatrix} \times 4$
maxpool, 2×2		
$(3 \times 3, 512) \times 3$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 640 \\ 3 \times 3, 640 \end{bmatrix} \times 4$
maxpool, 2×2		
$(3 \times 3, 512) \times 3$	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	
maxpool, 2×2		
fc-4096	avgpool	avgpool
fc-4096		
fc	fc	fc

C. Implementation details

This section provides all the additional implementation details for our experiments.

1) Model training: For the optimizer of SGD with momentum, the momentum factor is 0.9. For the RMSProp optimizer, the alpha parameter is 0.99. For the Adam optimizer, the beta parameter is $(0.9, 0.999)$. In the training process, the weight decay factor is set to $5\text{e}{-4}$, and the learning rate is initialized as 0.1, which is decayed by 0.2 every 30 epochs. For all training scenarios, we do not perform data augmentation during the training process.

2) Additional details in Section IV-B: We train VGG-16, ResNet-18, and WideResNet-28 on the source training set $\mathcal{S}$ and the concatenated set $\mathcal{S} \cup \mathcal{S}'$ with four different optimizers: SGD, SGD with momentum, RMSProp, and Adam. In the training procedure, the model is trained for 100 epochs. For

the model trained on the source training set $\mathcal{S}$, the batch size is set to 64, and the batch size is set to 128 for the model trained on the concatenated set to keep the number of iterations unchanged.

3) Additional details in Section IV-C: We employ the SGD optimizer with momentum to train VGG-16, ResNet-18, and WideResNet-28 on the source training set $\mathcal{S}$ and the concatenated set $\mathcal{S}\cup\mathcal{S}'$ with different source sample sizes $|\mathcal{S}|$ of $[2500, 5000, 7500, 10000, 12500, 15000, 17500, 20000, 22500, 25000]$, while the distraction set $\mathcal{S}'$ is invariant. The batch sizes in the different training processes follow the rule of $\frac{\text{batch size}}{\text{sample size}} = \frac{128}{50000}$ to keep the number of iterations unchanged.

4) Additional details in Section IV-D: We employ the optimizer of SGD with momentum to train VGG-16, ResNet-18, and WideResNet-28 on the source training set $\mathcal{S}$ and the concatenated set $\mathcal{S}\cup\mathcal{S}'$.

Distraction sample size. The sample size list of the distraction set $|\mathcal{S}'|$ is $[2500, 5000, 7500, 10000, 12500, 15000, 17500, 20000, 22500, 25000]$ and the source training set $\mathcal{S}$ is invariant *w.r.t.* the distraction sets with different sample sizes. Other details are the same as those in Section IV-D.

Label noise. For a specific label noise ratio $\alpha\%$, we randomly change the labels of $\alpha\%$ examples in the distraction set $\mathcal{S}'$. The list of label noise ratios is $[0.05, 0.1, 0.15, 0.2, 0.25]$. Other details are the same as those in Section IV-D.

Similarity between $\mathcal{S}$ and $\mathcal{S}'$. The experiments are conducted on CIFAR-10/100 other than Former/Latter CIFAR-10/100. Fifteen thousands images in the training sets of CIFAR-10/100 are randomly selected to constitute the distraction sets by adding color filters. For an RGB image img $= (C_1, C_2, C_3)$, where C_i denotes the i-th channel, the details of adding filters of level α to the image are presented as follows:

- **Red filter:** $(C_1, (1-\alpha)\,C_2, (1-\alpha)\,C_3)$;
- **Green filter:** $((1-\alpha)\,C_1, C_2, (1-\alpha)\,C_3)$
- **Red filter:** $((1-\alpha)\,C_1 + \alpha C_3, (1-\alpha)\,C_2 + \alpha C_3, C_3)$,

where α is the filter level and α is $[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$.

5) Additional details in Section IV-E: We generate $2,000$ training points and $10,000$ test points with the same normal vector $\mathbf{u}$ and generate 2000 distraction points with another normal vector $\mathbf{u}'$. We employ the SGD optimizer with momentum to train three-layer ReLU MLPs on the generated 150-dimensional data. In the training procedure, each model is trained for 20 epochs with a learning rate of 0.01. Each network f is repeatedly trained 10 times to estimate the random classifier $f_{\mathcal{A}}$ for calculating the distraction stability. In the experiments *w.r.t.* layer width, the list of layer widths is $[50, 100, 200, 500, 1000, 2000, 5000]$.

6) Additional details in Section V: CIFAR-10 and CIFAR-100 are divided into 10 subsets according to the categories $\{1,2,3,4,5,6,7,8,9,10\}$ and $\{1-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80, 81-90, 91-100\}$, respectively. The number of samples is fixed with $5,000$ for all ID estimation. The code of ID estimations with MLE can be downloaded here.

VIII. Proofs

To avoid technicalities, the measurability/integrability issues are ignored throughout this paper. Moreover, Fubini's theorem is assumed to be applicable for any integration w.r.t. multiple variables. In other words, the order of integrations is exchangeable.

A. Proof of Lemma 1

Our proof is an application of Theorem 7 in Nakada and Imaizumi [70], we first introduce some basic notions in the theorem.

Definition 7. *(Hölder space) Let $\gamma > 0$ be a degree of smoothness. For $f : [0,1]^n \to \mathbb{R}$, the Hölder norm is defined as*

$$\|f\|_{\mathcal{H}(\gamma,[0,1]^n)} := \max_{\alpha:\|\alpha\|_1<\lfloor\gamma\rfloor} \sup_{x\in[0,1]^n} |\partial^\alpha f(x)| + \max_{\alpha:\|\alpha\|_1=\lfloor\gamma\rfloor} \sup_{x,x'\in[0,1]^n, x\neq x'} \frac{|\partial^\alpha f(x) - \partial^\alpha f(x')|}{\|x-x'\|_\infty^{\gamma-\lfloor\gamma\rfloor}}. \quad (16)$$

Then, the Hölder space on $[0,1]^n$ is defined as

$$\mathcal{H}(\gamma,[0,1]^n) = \left\{ f \in C^{\lfloor\gamma\rfloor}([0,1]^n) \mid \|f\|_{\mathcal{H}(\gamma,[0,1]^n)} < \infty \right\}. \quad (17)$$

Also,

$$\mathcal{H}(\gamma,[0,1]^n, M) = \left\{ f \in \mathcal{H}(\gamma,[0,1]^n) \mid \|f\|_{\mathcal{H}(\gamma,[0,1]^n)} \leq M \right\} \quad (18)$$

denotes the M-radius closed ball in $\mathcal{H}(\gamma,[0,1]^n)$.

Definition 8. *(Labeling function) $f_0 : [0,1]^n \to \mathbb{R}$ is termed as the labeling function of the data distribution $\mathcal{D}$ if $y = f_0(\mathbf{x})$ for all $(\mathbf{x}, y) \in \mathcal{D}$.*

For the regression $\boldsymbol{\theta}^* \in \operatorname{argmin}_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^m (y_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i))^2$ *w.r.t.* the model $f : [0,1]^n \to \mathbb{R}$, the following lemma is proposed for the generalization error of the learned model $f_{\boldsymbol{\theta}^*}$

Lemma 2. *(Theorem 7 in [70]) If $f_0 \in \mathcal{H}(\gamma,[0,1]^n, M)$ and suppose ID$(\mathcal{D}_X) = d$. Then, there exists a constant C such that*

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(f_{\boldsymbol{\theta}^*}(\mathbf{x}) - y)^2 \right] \leq C m^{-2\gamma/(2\gamma+d)} (1+\log m)^2 \quad (19)$$

holds with probability at least $1 - 2\exp\left(-m^{d/(2\gamma+d)}\right)$ for any $m > N$ with a sufficiently large N.

Proof. We consider the binary classification where $f_0 : [0,1]^n \to \{0,1\}$. For the learned network $f_{\boldsymbol{\theta}^*} : [0,1]^n \to [0,1]^2$, we let $\widehat{f}_{\boldsymbol{\theta}^*}(\mathbf{x}) = \arg\max_i f^{(i)}_{\boldsymbol{\theta}^*}(\mathbf{x})$ to produce the prediction of $f_{\boldsymbol{\theta}^*}$. Then

$$\mathcal{R}_{\mathcal{D}}(\boldsymbol{\theta}^*) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbb{I}\left(y \neq f_{\boldsymbol{\theta}^*}(\mathbf{x})\right)\right] \quad (20)$$

$$= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\left(\widehat{f}_{\boldsymbol{\theta}^*}(\mathbf{x}) - f_0(\mathbf{x}) \right)^2 \right]. \quad (21)$$

With the zero training error assumption, we have $\boldsymbol{\theta}^* \in \operatorname{argmin}_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^m \left(f_0(\mathbf{x}_i) - \widehat{f}_{\boldsymbol{\theta}^*}(\mathbf{x}_i) \right)^2$. Therefore, according to Lemma 2, with the probability of at least $1 - 2\exp\left(-m^{d/(2\gamma+d)}\right)$ over a sample size of m, we have

$$\mathcal{R}_{\mathcal{D}}(\boldsymbol{\theta}^*) \leq C m^{-2\gamma/(2\gamma+d)} (1+\log m)^2. \quad (22)$$

Because the randomness of $\mathcal{R}_{\mathcal{D}}(\boldsymbol{\theta}^*)$ comes from the dataset $\mathcal{S} \sim \mathcal{D}^m$, the with the same of at least $1-2\exp\left(-m^{d/(2\gamma+d)}\right)$ over a sample size of m, we have

$$\mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m}\left[\mathcal{R}_{\mathcal{D}}\left(\mathcal{A}\left(\mathcal{S}\right)\right) \leq M\right] = \mathbb{P}_{\mathcal{S}\sim\mathcal{D}^m}\left[\mathcal{R}_{\mathcal{D}}\left(\boldsymbol{\theta}^*\right) \leq M\right], \tag{23}$$

where $Cm^{-2\gamma/(2\gamma+d)}(1+\log m)^2$, and the proof is completed. □

B. Proof of Proposition 1

Proof. From the definition of dataset-distraction stability, we have

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\frac{1}{2}\sum_{i=1}^{k}\left|f^{(i)}_{\mathcal{A}(\mathcal{S})}(\mathbf{x}) - f^{(i)}_{\mathcal{A}(\mathcal{S}\cup\mathcal{S}')}(\mathbf{x})\right|\right] \tag{24}$$

$$= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\left|f^{(y)}_{\mathcal{A}(\mathcal{S})}(\mathbf{x}) - f^{(y)}_{\mathcal{A}(\mathcal{S}\cup\mathcal{S}')}(\mathbf{x})\right|\right] = \beta \tag{25}$$

for the binary classification that $k = 2$. Recall that the LHS=$|\mathcal{R}_{\mathcal{D}}\left(\mathcal{A}\left(\mathcal{S}\cup\mathcal{S}'\right)\right) - \mathcal{R}_{\mathcal{D}}\left(\mathcal{A}\left(\mathcal{S}\right)\right)|$, then

$$\text{LHS} = \left|\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[f^{(y)}_{\mathcal{A}(\mathcal{S}\cup\mathcal{S}')}\right] - \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[f^{(y)}_{\mathcal{A}(\mathcal{S})}\right]\right| \tag{26}$$

$$= \left|\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[f^{(y)}_{\mathcal{A}(\mathcal{S}\cup\mathcal{S}')} - f^{(y)}_{\mathcal{A}(\mathcal{S})}\right]\right| \tag{27}$$

$$\leq \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\left|f^{(y)}_{\mathcal{A}(\mathcal{S}\cup\mathcal{S}')}(\mathbf{x}) - f^{(y)}_{\mathcal{A}(\mathcal{S})}(\mathbf{x})\right|\right] \tag{28}$$

$$= \beta \tag{29}$$

The proof is completed. □

C. Proof of Theorem 2

Proof. We start the proof from Eq. (11) such that $\mathcal{R}_{\mathcal{D}}(\mathcal{A}(\mathcal{S})) \leq \alpha_1\mathcal{R}_{\mathcal{D}_1}(\mathcal{A}(\mathcal{S}_1)) + \cdots + \alpha_T\mathcal{R}_{\mathcal{D}_T}(\mathcal{A}(\mathcal{S}_T)) + \alpha_1\beta_1 + \cdots + \alpha_T\beta_T$.

According to Lemma 1, for each $\mathcal{R}_{\mathcal{D}_i}(\mathcal{A}(\mathcal{S}_i))$ $(1 \leq i \leq T)$, we have

$$\mathcal{R}_{\mathcal{D}_i}(\mathcal{A}(\mathcal{S}_i)) \leq C|\mathcal{S}_i|^{-2\gamma/(2\gamma+d_i)}(1+\log|\mathcal{S}_i|)^2 \tag{30}$$

$$\leq C|\mathcal{S}_i|^{-2\gamma/(2\gamma+d_i)}(1+\log m)^2 \tag{31}$$

for any $m > N$ with a sufficiently large N. We note that $\mathcal{D}_i$ is the subset of $\mathcal{D}$, and it has the same labeling function as $\mathcal{D}$, *i.e.*, γ is the same bound as $\mathcal{D}$. If we assume $\min_i \frac{|\mathcal{S}_i|}{m} \geq \alpha > 0$ for all $\mathcal{S}$ that $|\mathcal{S}| = m > N$ and $\max_i d_i = d_{\max}$, then we have

$$\mathcal{R}_{\mathcal{D}_i}(\mathcal{A}(\mathcal{S}_i)) \leq C\left(\alpha m\right)^{-2\gamma/(2\gamma+d_i)}\left(1+\log m\right)^2 \tag{32}$$

$$\leq C\left(\alpha m\right)^{-2\gamma/(2\gamma+d_{\max})}\left(1+\log m\right)^2. \tag{33}$$

Then, with the probability of at last $\Pi_{i=1}^{T}1 - 2\exp\left(-m^{d_i/(2\gamma+d_i)}\right)$ over a sample size of m, we have

$$\mathcal{R}_{\mathcal{D}}(\mathcal{A}(\mathcal{S})) \tag{34}$$

$$\leq \alpha_1\mathcal{R}_{\mathcal{D}_1}(\mathcal{A}(\mathcal{S}_1)) + \cdots \alpha_t\mathcal{R}_{\mathcal{D}_T}(\mathcal{A}(\mathcal{S}_T)) + \alpha_1\beta_1 + \cdots \alpha_T\beta_T \tag{35}$$

$$\leq \sum_{i=1}^{T}\alpha_i C\left(\alpha m\right)^{-2\gamma/(2\gamma+d_{\max})}\left(1+\log m\right)^2 + \sum_{i=1}^{T}\alpha_i\beta_i \tag{36}$$

$$= C\left(\alpha m\right)^{-2\gamma/(2\gamma+d_{\max})}\left(1+\log m\right)^2 + \sum_{i=1}^{T}\alpha_i\beta_i, \tag{37}$$

and $C\alpha^{-2\gamma/(2\gamma+d_{\max})}$ is represented with C.

For the probability $\Pi_{i=1}^{T}1 - \exp\left(-m^{d_i/(2\gamma+d_i)}\right)$, we have

$$\Pi_{i=1}^{T}1 - 2\exp\left(-m^{d_i/(2\gamma+d_i)}\right) \tag{38}$$

$$\geq \left(1 - 2\exp\left(-m^{d_{\max}/(2\gamma+d_{\max})}\right)\right)^T \tag{39}$$

$$\geq 1 - 2T\exp\left(-m^{d_{\max}/(2\gamma+d_{\max})}\right). \tag{40}$$

With $d_{\max} < d$, the equality $-2T\exp\left(-m^{d_{\max}/(2\gamma+d_{\max})}\right) \geq -2\exp\left(-m^{d/(2\gamma+d)}\right)$ holds when $m \geq N$ for a sufficiently large N, and the proof is completed. □

D. Proof of Corollary 1

Proof. Recall Eq. 9 such that

$$\mathcal{R}_{\mathcal{D}}(\mathcal{A}(\mathcal{S})) = \alpha_1\mathcal{R}_{\mathcal{D}_1}(\mathcal{A}(\mathcal{S}_1\cup\cdots\cup\mathcal{S}_T)) + \cdots + \alpha_T\mathcal{R}_{\mathcal{D}_T}(\mathcal{A}(\mathcal{S}_1\cup\cdots\cup\mathcal{S}_T)). \tag{41}$$

According to the definition of expected risk, we have $\mathcal{R}_{\mathcal{D}_i} \leq 1$ for all $i \in \{1,\cdots,T\}$. Then, combined with Eq. 10 such that

$$\mathcal{R}_{\mathcal{D}_i}(\mathcal{A}(\mathcal{S}_1\cup\cdots\cup\mathcal{S}_T)) \leq \mathcal{R}_{\mathcal{D}_i}(\mathcal{A}(\mathcal{S}_i)) + \beta_i, \tag{42}$$

for all $t \in \{1,\cdots,T\}$, we can derive

$$\mathcal{R}_{\mathcal{D}}(\mathcal{A}(\mathcal{S})) = \alpha_1\mathcal{R}_{\mathcal{D}_1}(\mathcal{A}(\mathcal{S}_1\cup\cdots\cup\mathcal{S}_T)) + \cdots + \alpha_T\mathcal{R}_{\mathcal{D}_t}(\mathcal{A}(\mathcal{S}_1\cup\cdots\cup\mathcal{S}_T)) \tag{43}$$

$$\leq \sum_{i=t}^{T}\alpha_i\mathcal{R}_{\mathcal{D}_i}(\mathcal{A}(\mathcal{S}_1\cup\cdots\cup\mathcal{S}_T)) + \sum_{i=1}^{t}\alpha_i \tag{44}$$

$$\leq \sum_{i=t}^{T}\alpha_i\mathcal{R}_{\mathcal{D}_i}(\mathcal{A}(\mathcal{S}_i)) + \sum_{i=1}^{t}\alpha_i + \sum_{i=t}^{T}\alpha_i\beta_i. \tag{45}$$

By following the steps in the proof of Theorem 2 and $d_1 \geq \cdots \geq d_T$, we have

$$\sum_{i=t}^{T}\alpha_i\mathcal{R}_{\mathcal{D}_i}(\mathcal{A}(\mathcal{S}_i)) \leq C|\mathcal{S}_i|^{-2\gamma/(2\gamma+d_i)}(1+\log m)^2. \tag{46}$$

By substituting the this equation into Eq. 43,

$$\mathcal{R}_{\mathcal{D}}\left(\mathcal{A}(\mathcal{S})\right) \leq Cm^{-2\gamma/(2\gamma+d_t)}\left(1+\log m\right)^2 + \sum_{i=1}^{t}\alpha_i + \sum_{i=t}^{T}\alpha_i\beta_i. \tag{47}$$

Because the above inequality holds for all $t \in \{1,\cdots,T\}$, we have the following generalization bound:

$$\mathcal{R}_{\mathcal{D}}\left(\mathcal{A}(\mathcal{S})\right) \tag{48}$$

$$\leq \min_t\left\{Cm^{-2\gamma/(2\gamma+d_t)}\left(1+\log m\right)^2 + \sum_{i=1}^{t}\alpha_i + \sum_{i=t}^{T}\alpha_i\beta_i\right\} \tag{49}$$

and the proof is completed. □

IX. Conclusion

In this paper, we empirically and theoretically investigate the capability of attentive learning in neural networks, which is measured by a novel notion of dataset-distraction stability. A significant negative correlation between data-distraction stability and generalization performance is presented, and this correlation is not sensitive to the choice of distraction set. From the perspective of attentive learning, the learning problem can be decomposed into several simpler subtasks with lower dataset-distraction stability, which helps the network learn simple prediction rules. We further propose a tighter upper bound based on dataset-distraction stability and intrinsic dimension to support our findings. Therefore, attentive learning embodies the "divide and conquer" philosophy of learning these simpler subtasks separately with less interaction, which may be the source of the powerful generalizability in deep learning. Our research *w.r.t.* attentive learning not only helps to understand generalization in deep learning by explaining benign overfitting and the phenomenon of out-of-distribution distribution but also sheds light on novel algorithm designs such as active learning.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[2] 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*, 2016, pp. 770–778.

[3] G.-S. Xie, Z. Zhang, G. Liu, F. Zhu, L. Liu, L. Shao, and X. Li, "Generalized zero-shot learning with multiple graph adaptive generative networks," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 7, pp. 2903–2915, 2021.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[6] V. Mosin, I. Samenko, B. Kozlovskii, A. Tikhonov, and I. P. Yamshchikov, "Fine-tuning transformers: Vocabulary transfer," *Artificial Intelligence*, p. 103860, 2023.

[7] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[8] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[9] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[10] A. I. Maqueda, A. Loquercio, G. Gallego, N. García, and D. Scaramuzza, "Event-based vision meets deep learning on steering prediction for self-driving cars," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5419–5427.

[11] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *Journal of Field Robotics*, vol. 37, no. 3, pp. 362–386, 2020.

[12] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland *et al.*, "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, no. 7792, pp. 706–710, 2020.

[13] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[14] M. L. Baptista, K. Goebel, and E. M. Henriques, "Relation between prognostics predictor evaluation metrics and local interpretability shap values," *Artificial Intelligence*, vol. 306, p. 103667, 2022.

[15] I. Tiddi and S. Schlobach, "Knowledge graphs as tools for explainable machine learning: A survey," *Artificial Intelligence*, vol. 302, p. 103627, 2022.

[16] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.

[17] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, "Spectrally-normalized margin bounds for neural networks," *Advances in neural information processing systems*, vol. 30, 2017.

[18] B. Neyshabur, S. Bhojanapalli, and N. Srebro, "A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=Skz_WfbCZ

[19] H. Chen, F. He, S. Lei, and D. Tao, "Spectral complexity-scaled generalisation bound of complex-valued neural networks," *Artificial Intelligence*, p. 103951, 2023.

[20] K. Kawaguchi, "Deep learning without poor local minima," in *Advances in Neural Information Processing Systems*, 2016.

[21] H. Lu and K. Kawaguchi, "Depth creates no bad local minima," *arXiv preprint arXiv:1702.08580*, 2017.

[22] Y. Zhou and Y. Liang, "Critical points of neural networks: Analytical forms and landscape properties," in *International Conference on Learning Representations*, 2018.

[23] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, "The implicit bias of gradient descent on separable data," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 2822–2878, 2018.

[24] K. Lyu and J. Li, "Gradient descent maximizes the

margin of homogeneous neural networks," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SJeLIgBKPS

[25] Z. Ji and M. Telgarsky, "Directional convergence and alignment in deep learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 17 176–17 186. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/c76e4b2fa54f8506719a5c0dc14c2eb9-Paper.pdf

[26] D. Kalimeris, G. Kaplun, P. Nakkiran, B. Edelman, T. Yang, B. Barak, and H. Zhang, "Sgd on neural networks learns functions of increasing complexity," *Advances in Neural Information Processing Systems*, vol. 32, pp. 3496–3506, 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/b432f34c5a997c8e7c806a895ecc5e25-Paper.pdf

[27] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[28] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[29] Y. Lu, G. Lu, J. Li, Y. Xu, Z. Zhang, and D. Zhang, "Multiscale conditional regularization for convolutional neural networks," *IEEE Transactions on Cybernetics*, vol. 52, no. 1, pp. 444–458, 2020.

[30] Y. Lu, Z. Zhang, G. Lu, Y. Zhou, J. Li, and D. Zhang, "Addi-reg: A better generalization-optimization tradeoff regularization method for convolutional neural networks," *IEEE Transactions on Cybernetics*, vol. 52, no. 10, pp. 10 827–10 842, 2021.

[31] S. Wei, D. Murfet, M. Gong, H. Li, J. Gell-Redman, and T. Quella, "Deep learning is singular, and that's good," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2022.

[32] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=H1oyRlYgg

[33] F. He, T. Liu, and D. Tao, "Control batch size and learning rate to generalize well: Theoretical and empirical evidence," in *Advances in Neural Information Processing Systems*, 2019, pp. 1143–1152.

[34] F. He, B. Wang, and D. Tao, "Piecewise linear activations substantially shape the loss surfaces of neural networks," in *International Conference on Learning Representations*, 2020.

[35] L. Chizat and F. Bach, "Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss," in *Conference on Learning Theory*. PMLR, 2020, pp. 1305–1338.

[36] K. Lyu, Z. Li, R. Wang, and S. Arora, "Gradient descent on two-layer nets: Margin maximization and simplicity bias," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[37] D. Lin, R. Sun, and Z. Zhang, "On the landscape of one-hidden-layer sparse networks and beyond," *Artificial Intelligence*, vol. 309, p. 103739, 2022.

[38] S. Lei, "Understanding deep learning via large-scale systematic experiments," Master's thesis, The University of Sydney, 2021.

[39] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15 849–15 854, 2019.

[40] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, "Deep double descent: Where bigger models and more data hurt," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=B1g5sA4twr

[41] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, "On the spectral bias of neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5301–5310.

[42] Z.-Q. J. Xu, Y. Zhang, T. Luo, Y. Xiao, and Z. Ma, "Frequency principle: Fourier analysis sheds light on deep neural networks," *Communications in Computational Physics*, vol. 28, no. 5, pp. 1746–1767, 2020. [Online]. Available: http://global-sci.org/intro/article_detail/cicp/18395.html

[43] F. He, S. Lei, J. Ji, and D. Tao, "Neural networks behave as hash encoders: An empirical study," *arXiv preprint arXiv:2101.05490*, 2021.

[44] H. Wei, L. Tao, R. Xie, and B. An, "Open-set label noise can improve robustness against inherent label noise," *Advances in Neural Information Processing Systems*, vol. 34, pp. 7978–7992, 2021.

[45] G. Ortiz-Jimenez, A. Modas, S.-M. Moosavi, and P. Frossard, "Hold me tight! influence of discriminative features on deep network boundaries," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2935–2946, 2020.

[46] H. Shah, K. Tamuly, A. Raghunathan, P. Jain, and P. Netrapalli, "The pitfalls of simplicity bias in neural networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9573–9585, 2020.

[47] S. Lei, F. He, Y. Yuan, and D. Tao, "Understanding deep learning via decision boundary," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2023.

[48] D. L. Ruderman, "The statistics of natural images," *Network: computation in neural systems*, vol. 5, no. 4, p. 517, 1994.

[49] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[50] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[51] C. Fefferman, S. Mitter, and H. Narayanan, "Testing the manifold hypothesis," *Journal of the American Mathe-*

matical Society, vol. 29, no. 4, pp. 983–1049, 2016.

[52] S. Lei and D. Tao, "A comprehensive survey of dataset distillation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–17, 2023.

[53] P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein, "The intrinsic dimension of images and its impact on learning," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=XJk19XzGq2J

[54] K. M. Carter, R. Raich, and A. O. Hero III, "On local intrinsic dimension estimation and its applications," *IEEE Transactions on Signal Processing*, vol. 58, no. 2, pp. 650–663, 2009.

[55] S. Kpotufe, "k-nn regression adapts to local intrinsic dimension," *Advances in neural information processing systems*, vol. 24, 2011.

[56] M. Allegra, E. Facco, F. Denti, A. Laio, and A. Mira, "Data segmentation based on the local intrinsic dimension," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[57] P. Campadelli, E. Casiraghi, C. Ceruti, G. Lombardi, and A. Rozza, "Local intrinsic dimensionality based features for clustering," in *Image Analysis and Processing–ICIAP 2013: 17th International Conference, Naples, Italy, September 9-13, 2013. Proceedings, Part I 17*. Springer, 2013, pp. 41–50.

[58] K. Johnsson, C. Soneson, and M. Fontes, "Low bias local intrinsic dimension estimation from expected simplex skewness," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 1, pp. 196–202, 2014.

[59] D. Barbará and P. Chen, "Using the fractal dimension to cluster datasets," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000, pp. 260–264.

[60] J. D. Carroll and P. Arabie, "Multidimensional scaling," *Measurement, judgment and decision making*, pp. 179–250, 1998.

[61] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[62] I. Jolliffe, "Principal component analysis," *Encyclopedia of statistics in behavioral science*, 2005.

[63] B. Kégl, "Intrinsic dimension estimation using packing numbers," *Advances in neural information processing systems*, vol. 15, 2002.

[64] M. Fan, H. Qiao, and B. Zhang, "Intrinsic dimension estimation of manifolds by incising balls," *Pattern Recognition*, vol. 42, no. 5, pp. 780–787, 2009.

[65] R. Badii and A. Politi, "Hausdorff dimension and uniformity factor of strange attractors," *Physical review letters*, vol. 52, no. 19, p. 1661, 1984.

[66] E. Levina and P. Bickel, "Maximum likelihood estimation of intrinsic dimension," *Advances in neural information processing systems*, vol. 17, 2004.

[67] C. Ceruti, S. Bassis, A. Rozza, G. Lombardi, E. Casiraghi, and P. Campadelli, "Danco: An intrinsic dimensionality estimator exploiting angle and norm concentration," *Pattern recognition*, vol. 47, no. 8, pp. 2569–2581, 2014.

[68] E. Facco, M. d'Errico, A. Rodriguez, and A. Laio, "Estimating the intrinsic dimension of datasets by a minimal neighborhood information," *Scientific reports*, vol. 7, no. 1, p. 12140, 2017.

[69] H. Narayanan and S. Mitter, "Sample complexity of testing the manifold hypothesis," *Advances in neural information processing systems*, vol. 23, 2010.

[70] R. Nakada and M. Imaizumi, "Adaptive approximation and generalization of deep neural network with intrinsic dimensionality." *J. Mach. Learn. Res.*, vol. 21, pp. 174–1, 2020.

[71] F. Latorre, L. T. Dadi, P. Rolland, and V. Cevher, "The effect of the intrinsic dimension on the generalization of quadratic classifiers," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[72] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[73] O. Bousquet and A. Elisseeff, "Stability and generalization," *The Journal of Machine Learning Research*, vol. 2, pp. 499–526, 2002.

[74] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[75] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proceedings of the British Machine Vision Conference (BMVC)*, E. R. H. Richard C. Wilson and W. A. P. Smith, Eds. BMVA Press, September 2016, pp. 87.1–87.12. [Online]. Available: https://dx.doi.org/10.5244/C.30.87

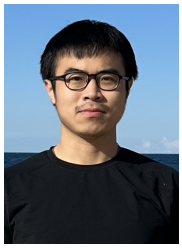
[76] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[77] G. Hinton, N. Srivastava, and K. Swersky, "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent," *Cited on*, vol. 14, no. 8, p. 2, 2012.

[78] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[79] J. P. Miller, R. Taori, A. Raghunathan, S. Sagawa, P. W. Koh, V. Shankar, P. Liang, Y. Carmon, and L. Schmidt, "Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7721–7735.

[80] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Tech. Rep., 2009.



Shiye Lei received MPhil in computer science from the University of Sydney in 2022. He is currently a PhD student in the School of Computer Science at the University of Sydney. His research interests include data-centric AI and knowledge extraction of large-scale datasets.

Fengxiang He is a Lecturer at Artificial Intelligence and its Applications Institute, School of Informatics, University of Edinburgh, and an Affiliate of Edinburgh Future Institute. He received BSc in statistics from University of Science and Technology of China, MPhil and PhD in computer science from the University of Sydney. His research interest is in trustworthy AI, particularly deep learning theory, theory of decentralised learning, learning theory in game-theoretical problems, and their applications. He is an Area Chair of UAI, AISTATS, and ACML.

Haowen Chen received Bachelor in Science from University of Hong Kong in 2022. She is currently a MSc student in the Department of Mathematics at Swiss Federal Institute of Technology in Zürich (ETH Zürich). Her research interests include stochastic analysis, approximation theory of deep neural networks, and statistical learning theory.

Dacheng Tao (Fellow, IEEE) holds the Peter Nicol Russell Chair in School of Computer Science at the University of Sydney. He mainly applies statistics and mathematics to artificial intelligence and data science, and his research is detailed in one monograph and over 200 publications in prestigious journals and proceedings at leading conferences, with best paper awards, best student paper awards, and test-of-time awards. His publications have been cited over 112K times and he has an h-index 160+ in Google Scholar. He received the 2015 and 2020 Australian Eureka Prize, the 2018 IEEE ICDM Research Contributions Award, and the 2021 IEEE Computer Society McCluskey Technical Achievement Award. He is a Fellow of the Australian Academy of Science, AAAS, ACM and IEEE.