

Received April 22, 2020, accepted May 5, 2020, date of publication May 8, 2020, date of current version May 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2993304

# Towards Adversarial Robustness via Feature Matching

ZHUORONG LI<sup>1</sup>, CHAO FENG<sup>2</sup>, JIANWEI ZHENG<sup>3</sup>, MINGHUI WU<sup>1</sup>,  
AND HONGCHUAN YU<sup>4</sup>, (Member, IEEE)

<sup>1</sup>School of Computer and Computing Science, Zhejiang University City College, Hangzhou 310015, China

<sup>2</sup>College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

<sup>3</sup>College of Computer Science and Engineering, Zhejiang University of Technology, Hangzhou 310014, China

<sup>4</sup>National Centre for Computer Animation, Bournemouth University, Poole BH12 5BB, U.K.

Corresponding author: Minghui Wu (mhwu@zucc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61602413, the Natural Science Foundation of Zhejiang Province under Grant LY19F030016, and the EU H2020 project-AniAge under Grant 691215.

**ABSTRACT** Image classification systems are known to be vulnerable to adversarial attacks, which are imperceptibly perturbed but lead to spectacularly disgraceful classification. Adversarial training is one of the most effective defenses for improving the robustness of classifiers. We introduce an enhanced adversarial training approach in this work. Motivated by human's consistently accurate perception of surroundings, we explore the artificial attention of deep neural networks in the context of adversarial classification. We begin with an empirical analysis of how the attention of artificial systems will change as the model undergoes adversarial attacks. Observation is that the class-specific attention gets diverted and subsequently induces wrong prediction. To that end, we propose a regularizer encouraging the consistency in the artificial attention on the clean image and its adversarial counterpart. Our method shows improved empirical robustness over the state-of-the-art, secures 55.74% adversarial accuracy on CIFAR-10 with perturbation budget of 8/255 under the challenging untargeted attack in white-box settings. Further evaluations on CIFAR-100 also show our potential for a desirable boost in adversarial robustness for deep neural networks. Code and trained models of our work are available at: <https://github.com/lizhuorong/Towards-Adversarial-Robustness-via-Feature-matching>


**INDEX TERMS** Bio-inspired explanations, deep learning, defense, adversarial attack, learning representations.

## I. INTRODUCTION

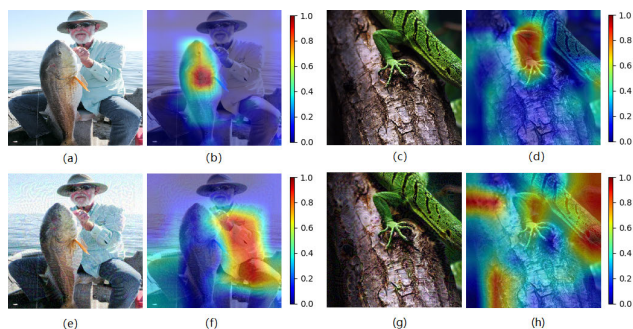
Whereas deep neural networks perform a variety of computer vision tasks with superior accuracies, their performance spectacularly degrades under ubiquitous threat of the adversarial attacks [1]. In the context of image classification, adversarial attacks are crafted from natural image with imperceptible perturbation to induce erroneous predictions [2]–[4]. Even worse, the attacked classifier outputs the incorrect prediction with surprisingly high confidence. Serious concerns are raised when the deep neural networks are applied to real-world applications, especially on reliability and security systems [5]–[8]. This problem has garnered enormous attention and encourages high activity on defense

methods [9]–[13], which can be roughly categorized into three catalogs: using network add-on, changing network architecture and adversarial training [14]. Among them, adversarial training offers the state-of-the-art robustness. In this paper, we focus on exploring an enhanced adversarial training method.

While deep neural networks are fragile to such subtle perturbations, we humans are still able to make correct prediction on these deceiving images. Motivated by the outstanding reliability and efficiency of human brain, we devise a method learning from how the brain classifies. To effectively defend the adversarial attack, we first investigate what happens to the victim models as they undergo attacks. We leverage a bio-inspired technique, Class Activation Mapping (CAM) [15], [16], for our empirical analysis. Motivation is that, being statistically demonstrated highly correlative with

The associate editor coordinating the review of this manuscript and approving it for publication was Carlos M. M Travieso-González .

human attention [17], CAM clearly indicates the discriminative regions used by the network to recognize classes. As shown in Fig.1, the class activation maps of original images (see Fig.1(a) and (c)) are highly class-discriminative, with visual explanation exclusively highlight the regions of “barracouta” and “green lizard” separately. When apply adversarial perturbations to the original images, though they are too subtle to detect (see Fig.1(e) and (g)), the alternations on class activation maps are substantial. Fig.1(f) shows that the that the attention of the network on the true object, “barracouta”, has been diverted to other objects and thus making wrong prediction, “sea snake”. Also, in Fig.1(h), the region of interest, i.e. where the class-specific features of “green lizard” lie, is dispersed over the background under attacks, leading to a misclassification as “common newt”.



**FIGURE 1.** Imperceptible perturbation in pixel space can result in significant alteration on class activation maps. Here shows the class activation maps in the penultimate layer of VGG16 corresponding to natural images (top) and adversarial images (bottom) respectively. The adversarial images are generated by PGD [18] with perturbation budget 8/255. The adversarial examples are incorrectly predicted as “sea snake” and “common newt”, while the true labels are “barracouta” and “green lizard”. Note that red indicates regions with high score for the class.

The empirical study above shows that, though the perturbation is subtle in pixel space, it can be easily detected through the significant attention shift in feature space. Note that a robust classifier is supposed to be insensitive to adversarial noise thus holding the original prediction. Through the lens of class activation maps, the original class-discriminative attention will stay where it was with unchanged confidence. Building on this intuition, we introduce a regularization term to penalize the alternations on class activation maps, so that it becomes difficult to induce the wrong prediction by adding an adversarial noise. In the framework of achieving robustness by feature matching, we also investigate alternative feature matching operations.

Model trained by the proposed method substantially improves the state-of-the-art adversarial robustness under a wide range of strong attacks in white-box settings, on standard benchmark dataset CIFAR-10 and even the challenging CIFAR-100.

To summarize, our main contributions are as below:

- 1) We present an enhanced defense which encourages the consistency in class-discriminative feature between the clean image and its adversarial counterpart by matching

the class activation maps. We also provide a variants of feature matching operations to seek the defense that exhibits the best robustness.

- 2) We achieve new state-of-the-art adversarial accuracy on CIFAR-10 with perturbation budget of 8/255 under untargeted attack in the highly challenging white-box settings. Specifically, we get 51.54%, 55.74% and 52.95% adversarial accuracy under strong attacks named C&W(30), PGD (7) and PGD (20), separately.
- 3) For further evaluation on the much challenging dataset, we implement several baselines from the literature on CIFAR-100 and conduct comparisons on it. Our method exceeds the state-of-the-art methods under a wide range of attacks.

## II. BACKGROUND AND RELATED WORK

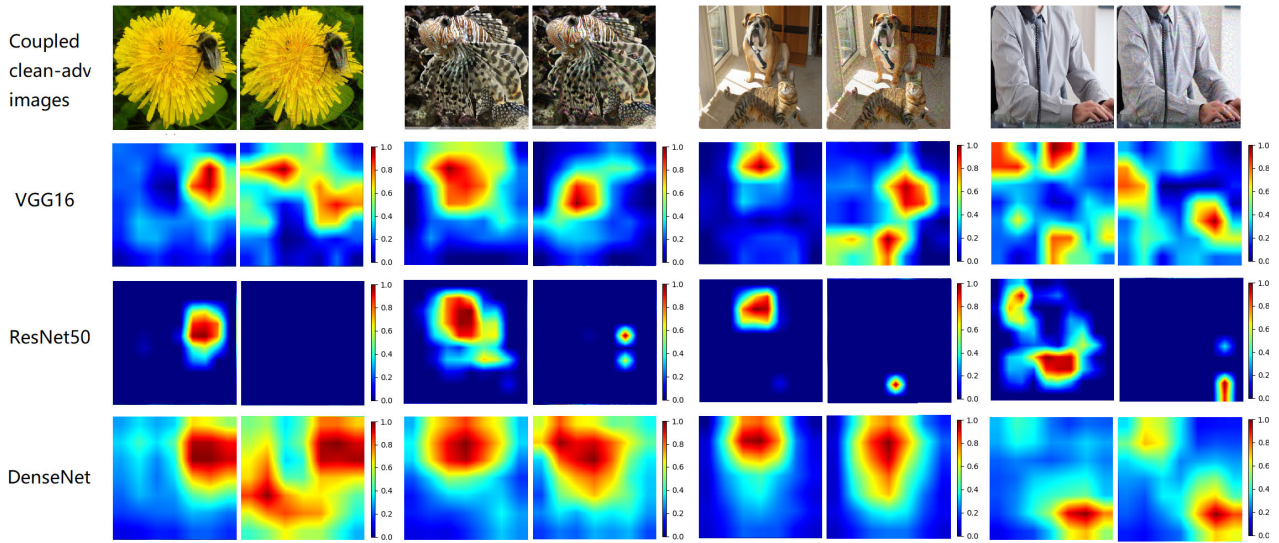
Consider a standard classification task. Given an example  $x \in \mathbb{R}^d$  and the corresponding label  $y \in [k]$  that drawn from an underlying data distribution  $\mathcal{D}$ , as well as the predefined loss function  $L$ , e.g., the widely used cross-entropy loss in image classification task. The goal is to find parameters  $\theta$  that minimize the population risk  $\mathbb{E}_{(x,y) \sim \mathcal{D}}[L(x; y, \theta)]$ , which is also known as the Empirical Risk Minimization (ERM).

Though models that trained by ERM work well on the holdout test data, they degrade spectacularly under the adversarial attacks due to the induced distribution shift [2], [9]. Many efforts have been devoted to securing the model against adversarial examples [10], [18]–[21]. As one of the most strongest defenses, adversarial training effectively alleviates the issues raised by the distribution shift. It adapts the ERM paradigm to adversarial images, towards improving robustness. In practice, the adversarial examples are crafted from the original training data on-the-fly and fed into the model as inputs during training. Thereby, the distribution shift can be effectively rectified.

There has been significant progress in developing stronger adversarial training methods. A notable method has been proposed by Madry *et al.* [18], which casts the defense into an optimization task and solves it reliably. Specifically, it seeks the parameter  $\delta^*$  that optimizes the following min-max problem:

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\max_{\delta \in S} L(x; y, \theta)] \quad (1)$$

where  $\delta$  is the perturbation that subjects to  $l_p$ -norm budget as  $\|\delta\|_p \leq \epsilon$  and  $S$  is the set of allowed perturbations. Whereas the inner maximization corresponds to the generation of attacks, the outer optimization is to defend against these attacks via minimizing the loss induced by the adversarial attacks. Madry *et al.* [18] also suggests that universal robustness can be developed through adopting the first-order adversary, projected gradient descent (PGD) attack, for adversarial training. For simplicity, we abbreviate this method to AT-PGD hereafter for simplicity. The resulting models of AT-PGD achieved the first empirical robustness on popular CIFAR-10 dataset and thus being one of the most popular baselines on adversarial training. While it effectively



**FIGURE 2.** Examples of the paired class-specific feature maps on the coupled clean-adversarial images across various DNNs. From left to right, clean images with true classes “bee”, “lionfish”, “bull mastiff” and “lab coat”, are crafted into adversarial images to induce incorrect prediction, namely “sulphur butterfly”, “whiptail”, “panthera tigris” and “box turtle”. Though the patterns of attention shift might be distinct, the discrepancy of feature maps induced by the adversarial perturbations can be consistently observed across different types of DNNs.

improves the robustness over the previous method, it disregards the connections between the natural examples and the corresponding perturbed versions [22].

Building on AT-PGD [18], Kannan *et al.* [23] recently propose Adversarial Logit Pairing (ALP), which improves the performance by an extra constraint. Specifically, apart from the same predicted class enforced by AT-PGD [18], they also encourage the similarity in the logits vector between the natural examples and the corresponding crafted versions:

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} [L(x; y, \theta) + \lambda D(f(x; \theta), f(x + \delta^*; \theta))] \quad (2)$$

The first term guarantees the performance on natural examples while the additional term is designed to guide the matching of the logits vector  $f(x; \theta)$  and  $f(x + \delta^*; \theta)$ . Here,  $\lambda$  is a hyperparameter and  $D$  is a measure for distance, e.g.,  $\ell_2$  loss. They suggest that the logit pairing is beneficial as it utilizes the inter-correlation of the example pair. Nevertheless, ALP has not been verified under the untargeted attacks in the literature. Engstrom *et al.* [24] suggests that models in [23] are only trained to defend against targeted attacks (e.g. with intentional classes or random target classes), thus being weaker than the ones trained to resist against untargeted attacks.

Concurrent to our work, Ref. [25] frames the adversarial training as feature denoising, which highlights scaling the adversarial training to ImageNet, requiring modification to the network architecture with additional denoising blocks. As shown in Fig.2, attention shifts can be consistently induced by adversarial perturbations across various Deep Neural Networks (DNNs). Motivated by this observation, we alternatively elaborate our method to be applicable to any pretrained backbone models. The main difference

between our work and feature distillation firstly introduced by Hinton *et al.* [26] is that we keep the same network architecture in training and further apply the distilled knowledge as a defense. Orthogonal to our work, Zhang and Wang [22] introduces Optimal Transport solver to improve robustness over the Kullback-Leibler (KL) or Jensen-Shannon (JS) based methods. In addition, model ensemble [11], [27] and unsupervised scheme [22], [28] also enable the robustness improvement. It might be possible to achieve further improvement by appropriately integrate these techniques in future work, but it is beyond the scope of this paper.

In this work, we present a regularization technique to enhance the defense against universal adversarial attacks. Motivated by the success of AT-PGD [18], we use it as the underlying basis for our approach. In contrast to it, we take advantage of the inter-relationship between the original sample and its counterpart in addition to the shared label. This draws our method closely related to the state-of-the-art method, ALP [23], which uses logits vector as the feature and highlighted the logits pairing for the defense against targeted attacks. Our method mainly differs in two perspectives, the choice of feature for matching and the attacks we defend against. As for the feature for matching, we adopt the gradient-weighted class activation of the penultimate layers as meaningful feature. It enables the fully preservation of the spatial and class-discriminative information that crucial for training a good classifier, as justified in extensive comparative experiments in Sec.IV. With respect to the attacks, we aim at defending against untargeted attacks rather than targeted attacks that used in ALP, as suggested in [24] that a defense robust to untargeted attacks is stronger than the one only robust to targeted attacks.

### III. PROPOSED METHOD

#### A. FEATURE MATCHING

In this work, we propose to leverage a feature matching operation in the form of an additional regularizer to the standard adversarial training.

##### 1) LAYERS TO SELECT

Prior work [23] proposes to match the logits vector, that is, the activation after the fully connected layer. Alternatively, we encourage the consistency in the penultimate layer (i.e., the nearest convolution layer to the fully connected layer, which is also named pre-dense layer) to suppress the alteration in class prediction. Our motivation for the selected layer is that, the spatial information can be retained in convolution layers, while will get lost in the dense layer, as all these spatial features will be flattened to output a vector of logits. In addition, higher-level visual components can be captured as the layer goes deeper [29], [30]. Therefore, the last convolution layer is expected to be considerably expressive. We provide quantitative experiments in Sec-IV.D to justify our layer selection for feature matching.

##### 2) FEATURES FOR MATCHING

However, the naïve implement by directly matching the feature maps in penultimate layer between the images is not effective, as the spatial structure maintained in higher layers might be not class-specific. A simple alternative to the single-layer scheme is matching the feature maps of multilayers. By means of comparative experiments, we found that it still suffers from this problem at a reduced level, which will be presented in Sec.IV-D. We speculate the comparability of these two schemes is that, the increment information of multilayers over one deepest layer are those preserved in the shallower layers, e.g., colors and details of images. These features might be important for image reconstruction but not significantly beneficial for our purpose. In addition, the perturbation in feature space is easy to observe but difficult to measure [25], which poses another challenge for feature matching.

To address above issues, we propose to leverage a much class-discriminative and easier quantitated feature for our feature matching. As a robust classifier is supposed to be insensitive to noise and will remain the original predictions. Our institution is that, the class-specific attention, which corresponds to the region of interest, will stay where it was with confidence. Leveraging the visual explanations of neural networks [15], [16], we use the combination of  $k$  feature maps in penultimate layer  $A$  followed by the Global Average Pooling  $g$  and ReLU activation as the features to match:

$$\varphi_{CAM}^c = \text{ReLU}\left(\sum_k g\left(\frac{\partial y^c}{\partial A^k}\right)\right) \quad (3)$$

where the gradients  $(\frac{\partial y^c}{\partial A^k})$  are back propagated by the true class  $c$ . Gradients are set to zero for all categories except the ground-true class, on purpose of class discrimination.

ReLU activations used here is to only highlight the features with positive impact on the attention [16]. Therefore, we can encourage the similarity in truly meaningful embedding of the paired examples by the suppression of the attention shift induced by perturbations.

Note that we do not use the CAM technique [15] and its extension [16] in an exact way as in the literature. Concretely, we deprecate the normalization when apply the class-specific attention to adversarial feature matching. This is because the the normalization to  $[0,1]$  in CAM is for visualization purposes, but it will lead to inessential scaling that undesirably undermines the feature mapping.

##### 3) FEATURE DISTANCE

As recommended by Johnson *et al.* [31], we adopt the feature reconstruction loss as a measure for our matching operation. The feature distance between two images  $x$  and  $\hat{x}$  is defined as the distance between representation  $\varphi_j(\cdot)$  of images in layer of the network:

$$D_{feat}^{\varphi_j}(x, \hat{x}) = \frac{1}{C_j H_j W_j} \|\varphi_j(x), \varphi_j(\hat{x})\|_2^2 \quad (4)$$

where  $C_j \times H_j \times W_j$  is the shape of feature map of layer  $j$ .

##### 4) ADVERSARIAL FEATURE MATCHING

Given a clean image  $x$ , the proposed adversarial feature matching is implemented by minimizing the distance between the class-specific activation of the original image  $x$  and the corresponding adversarial image:

$$D_{CAM}(\varphi_{CAM}^c(\theta, x), \varphi_{CAM}^c(\theta, x + \delta^*)) \quad (5)$$

### B. ADVERSARIAL TRAINING WITH FEATURE MATCHING REGULARIZER

Trained by the cross-entropy loss only, standard classifiers might be able to fit the training distribution, but it is prone to have undesirable behavior off the data manifold. ALP [23] achieves better robustness by using adversarial examples as additional input, which guides better behaviors on a larger region. ALP further enforces a loss term for better understanding of data.

Our empirical analysis above suggests that the class-specific activation can well represent the imperceptible and offensive perturbation in the form of attention shift in the feature space (see Fig.1). Therefore, we exploit the feature matching operation as a regularizer for enhanced adversarial training, with the total loss function as follow:

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} [L(x; y, \theta, \delta^*) + \lambda_1 D_{CAM}(\varphi_{CAM}^c(\theta, x), \varphi_{CAM}^c(\theta, x + \delta^*)) + \lambda_2 L_{norm}] \quad (6)$$

Here the first term refers to Eq.(2) as the experiments show that our regularizer works better when used in conjunction with Eq.(2) than with Eq.(1). The second term signifies the proposed regularizer, and the last term  $L_{norm} = \|\varphi_{CAM}^c(\theta, x)\|_2 + \|\varphi_{CAM}^c(\theta, x + \delta^*)\|_2$  is the  $\ell_2$  norm decay that

widely used in image processing tasks [32]. Coefficients  $\lambda_1$  and  $\lambda_2$  are weights for the trade-off among loss terms.

We compare all the variants described above in Sec.IV-D, with results showing that our method attains substantial outperformance over the baseline models. We hypothesize that the proposed regularization works well as it essentially provides an extra prior to guide the model towards better representation of data.

## IV. EXPERIMENTS AND RESULTS

### A. EXPERIMENTS SETTINGS

Following common protocols [18] for evaluating the adversarial training models, we consider the untargeted attacks, since a defense robust to untargeted adversarial attacks is stronger than the one only robust to targeted attacks [24]. In this work, we perform the evaluation under the white-box settings, where the adversary has access to the parameters of model that to be attacked.

We perform extensive experiments on two benchmark datasets, CIFAR-10 and CIFAR-100 [33]. The former is widely used in adversarial training literature [18], [22], [34], [35] and the latter is more challenging as the number of training images per class is only one-tenth of that in CIFAR-10. The baseline methods we compare against are: (1) standard training with clean images only (Standard), (2) the min-max optimization based adversarial training (AT-PGD) [18], which is one of the most robust defense approach, and the adversarial logit pairing (ALP) [23], which is currently the state-of-the-art adversarial training technique.

To be comparable with baselines, we train the models against the  $\ell_\infty$ -bounded adversaries, which are generated by the PGD optimization with size of 2 for 7 steps, and the total perturbation budget  $\epsilon = 8$ . Data augmentations including random flips, crops and per image normalization are employed. We adopt the wider variant of ResNet as the network architecture, which is with larger capacity than the original one and benefits a lot especially when the adversaries are imperceptible.

As for evaluation, in order to have a close measurement of the true robustness, we test all the models against a variety of threat models, including FGSM [4], C&W [36], and PGD attacks in white-box settings. The metric we measure is the adversarial accuracy, i.e., the percentage of correctly classified images on the test set that are perturbed by the threat model. Constrained by the expensive computation cost, we limit to compare all models on the first 1K test images.

### B. EVALUATING THE EMPIRICAL ROBUSTNESS

#### 1) RESISTANCE FOR DIFFERENT UNTARGETED ATTACKS

CIFAR-10 consists of 50K training images and 10K test images in 10 classes. To be comparable, we follow the same hyperparameters settings as the widely used baselines [18]. Details can be found in our experiment settings. We summarize the adversarial classification accuracy on the original images (Clean) and various attacks in Table 1. It shows that

**TABLE 1. Classification accuracy on CIFAR-10. Best results under each attack are in bold to show the best performance while lowest bound for each method is underlined, depicting the most threatening attack [40]. The proposed method performs the best under a wide range of challenging untargeted attacks.**

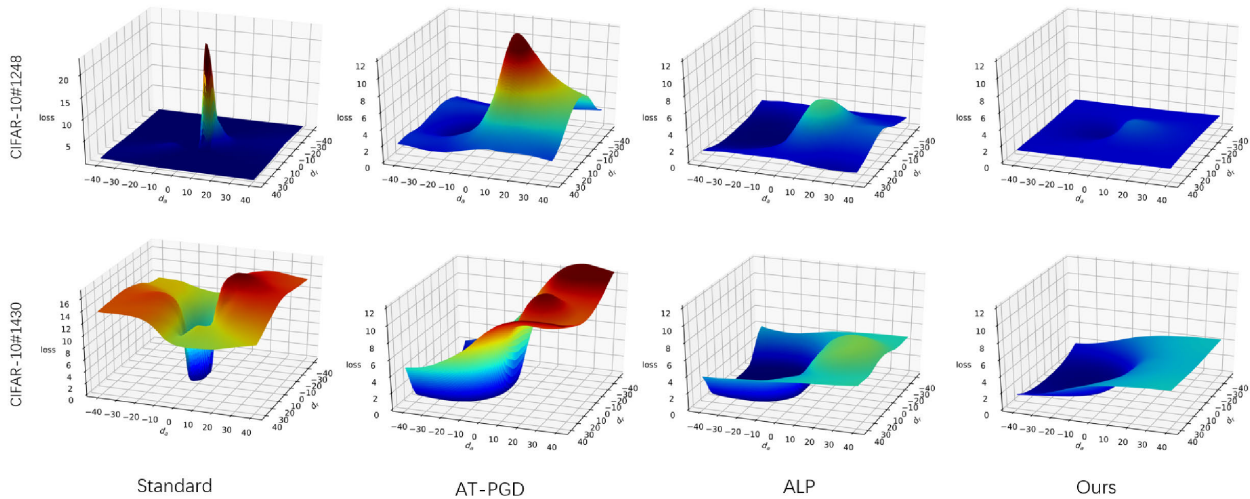
CIFAR-10( $\ell_\infty=8/255$ )					
Attacks (steps)	Clean	FGSM (1)	C&W (30)	PGD (7)	PGD (20)
Attack Strength		Weak	Strong		
Standard	<b>95.01</b>	13.35	0.00	0.00	0.00
AT-PGD	87.25	56.22	46.70	49.70	<u>45.87</u>
ALP	86.52	60.57	46.80	52.32	46.28
Ours	85.21	<b>60.10</b>	<b>51.54</b>	<b>55.74</b>	<b>52.95</b>

Standard model fails under adversarial attacks. AT-PGD and ALP significantly improves the robustness over Standard model by achieving 45.87% and 46.28% accuracy separately under the 20-step PGD attack. Model trained with the proposed method achieves desirable outperformance with clear margins under the strong attacks, e.g., C&W attack and PGD attacks. When applied weaker attack, e.g. one step FGSM, our method is still able to tie with the state-of-the-art. Note that we get decrease of accuracy on clean images, but this is also observed in all existing defense models [34], [37] due to an inherent trade-off between the robustness to adversarial attacks and its standard accuracy [38]. This is caused by the fragile correlation between the learned features of the classifier and the label [39]. To achieve almost perfect performance, standard classifiers tend to take advantage of any feature that useful for classification, even the feature that can only provide weak information. However, the paradigm of adversarially trained model is different as it mitigates adversarial examples by rejecting these non-robust features, which are only slightly correlated to label and easily manipulated. Therefore, when the defended model applies on clean images, the standard classification performance will inevitably degrade as some features have been discarded. Nevertheless, we believe such slight dip will be outweighed by the considerable gains in adversarial robustness.

CIFAR-100 is of the same number of training images and also the test images as CIFAR-10, but it belongs to 100 classes. Much less image per class makes it more challenging than CIFAR-10, which can be illustrated by the lower accuracies in Table 2 as compared to Table 1. Our method still outperforms the baselines no matter under weak attacks or

**TABLE 2. Classification accuracy comparison under different untargeted attacks on CIFAR100. The proposed method performs the best under a wide range of untargeted attacks.**

CIFAR-100( $\ell_\infty=8/255$ )					
Attacks (steps)	Clean	FGSM (1)	C&W (10)	PGD (10)	PGD (20)
Attack Strength		Weak	Strong		
Standard	<b>66.50</b>	5.07	0.00	0.00	0.00
AT-PGD	61.14	29.24	24.39	24.82	24.09
ALP	66.04	31.83	26.01	27.50	26.60
Ours	59.51	<b>34.49</b>	<b>28.00</b>	<b>31.68</b>	<b>31.11</b>



**FIGURE 3.** Comparison of loss surface for different models in the vicinity of natural examples 1248 (top) and 1430 (bottom) of CIFAR-10 validation set, with the true label of “automobile” and “horse” respectively. The loss is represented by z-axis, which is varying with the adversarial perturbation ( $da = \text{sign}(\nabla_x f(x))$ ) and the Rademacher vector ( $dr \sim \text{Rademacher}(0.5)$ ).

strong attacks, which further demonstrates the effectiveness of the proposed method.

Fig.3 plots the loss landscapes of different models as another comparison. The loss surface of the undefended model that trained by the standard cross-entropy loss only, is highly bumpy, see Fig.3 (a). While AT-PGD gets lower loss, the loss surface is still highly non-linear. In other words, ALP might be able to improve the robustness, but it still suffers the same problem at a reduced level. Noticeably, Fig.3 (d) shows the substantial improvement by the proposed method, achieving further loss reduction and flattened landscape.

## 2) RESISTANCE FOR ATTACKS OF DIFFERENT STRENGTH

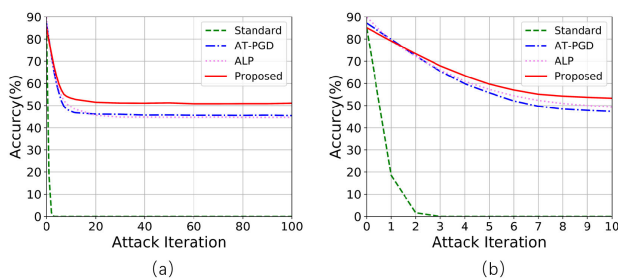
We further evaluate the resistance of our method against attacks of different strengths. Madry *et al.* [18] suggests that being robust against PGD implies resistance against many other first order attacks. Therefore, we set PGD as the adversary in this experiment and vary the iterations of it, indicating different strengths of attacks. Fig.4 shows the model

robustness under PGD attacks that with different strength. It can be observed that except for the failure of the undefended Standard model at an early iteration, all defended models can sustain a stable accuracy along with increasing attack iterations. Note that the proposed method consistently outperforms all the baselines by a clear margin after the attack convergence. It suggests that the proposed model is fairly strong against the attacks across a wide range of iterations.

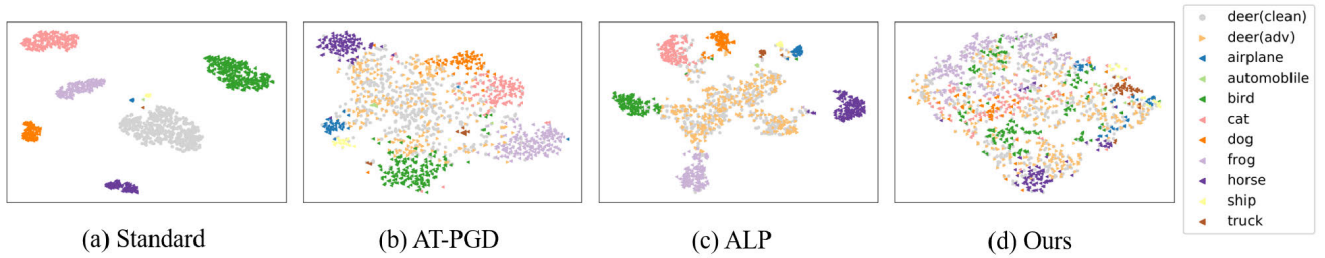
## C. QUALITATIVE EVALUATION

In addition to the quantitative evaluation above, we also conduct the qualitative analysis by visualizing the latent representation. Our insight is that a robust model can push the cluster of adversarial examples closer to the true class, so that the adversarial images and their clean counterparts will be clustered together as they belong to the same true class. Fig.5 visualizes the penultimate fully connected layer of the proposed model and the baseline models. Unsurprisingly, the undefended model (a) perfectly groups all the clean images into the same cluster. However, it separates the adversarial examples into distinct groups that far away from the clean images. In other words, the standard trained model misclassifies almost all the adversarial images into false class. This observation is consistent with the results of Standard in Table 1. Though AT-PGD and ALP can draw the adversarial examples and clean images closer, but several distinct clusters can still be observed in Fig.5(b) and (c). On the contrary, our method draws all the adversarial examples closer without distinct clusters. In other words, the proposed defense is strong enough that it is difficult to generate adversarial examples to fool the models trained with our method.

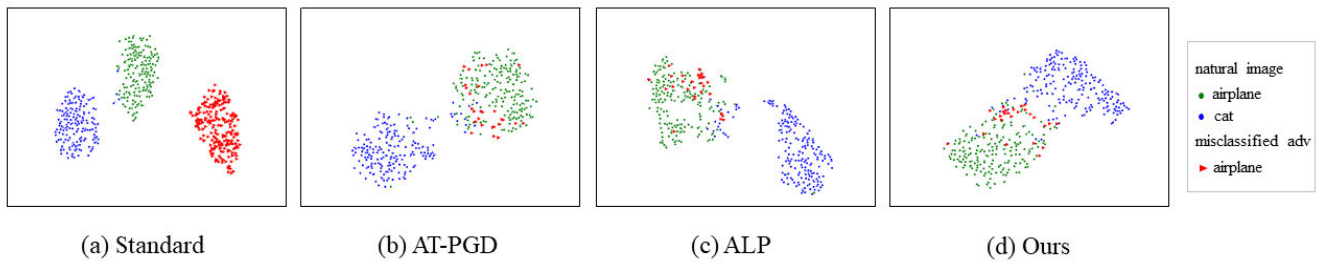
We further present the separation of clean and adversarial images to illustrate how the images are tempted



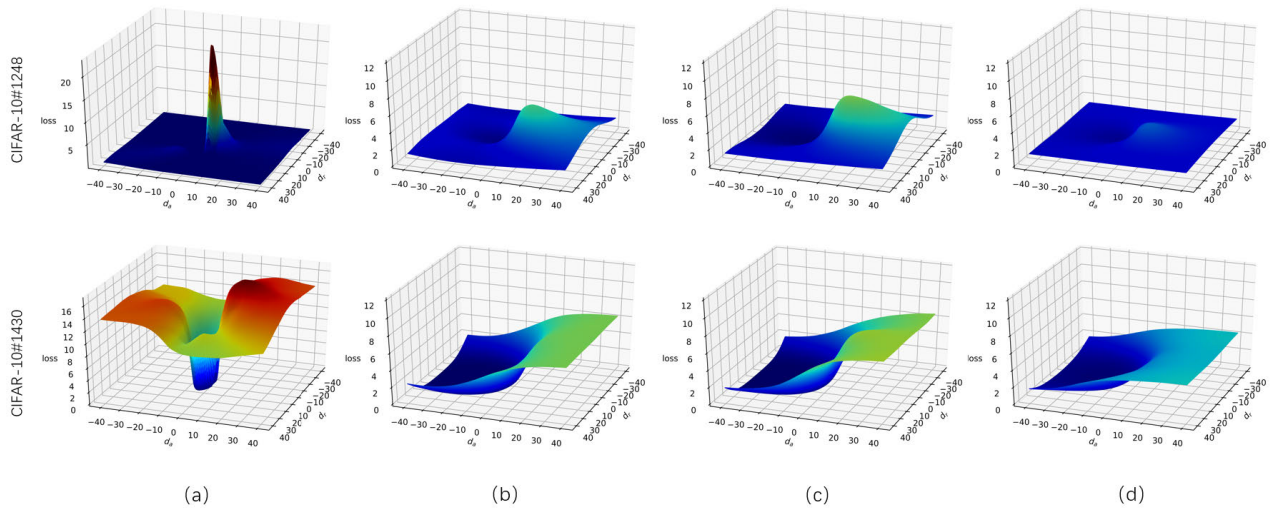
**FIGURE 4.** Model robustness under PGD attacks with different iterations. (a) Results from 1 to 100 iterations. (b) Zooms in the results within 10 iteration. Models are trained with attacks with perturbation budget of 8/255 and attack iteration of 7. Results show that the proposed model is secured under a wide range of strength of threat.



**FIGURE 5.** t-SNE visualizations of latent representation under adversarial attacks. Adversarial examples are crafted from 1000 clean test images of CIFAR-10. Clean images are denoted by dots while the adversarial examples are represented by colored triangles. Among them, triangle that towards the right denotes the correctly classified adversarial examples, and the other ones denote the misclassified images. Noticeable is that the undefended model (see (a)) separates the adversarial examples into distinct groups with clear margins. (b) AT-PGD and (c) ALP are more robust than (a) Standard but are still less than ideal, as clusters can still be distinguished. In contrast, the proposed model (d) gathers all the samples together as they actually belong to the same class, demonstrating its robustness and superiority over the baselines.



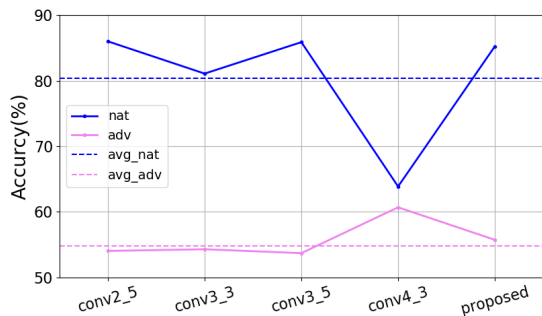
**FIGURE 6.** The separation of adversarial images from the natural samples. 200 random images sampled from “airplane” and “cat” of CIFAR-10 are denoted by green and blue dots separately. Adversarial examples crafted from the natural images but misclassified into false class are denoted by red triangles. It is shown that for (a) standard classifier, the adversarial images are far from the true cluster. (b) AT-PGD and (c) ALP show improved robustness with less mistakenly classified images when compared with (a). Notice that the proposed method (d) further mitigates the malicious example by drawing them closer to the true class “cat”.



**FIGURE 7.** Comparison of loss surface for different feature matching operations around data points 1248 (top) and 1430 (bottom) of CIFAR-10 validation set: (a) Standard; (b) multilayers; (c) single layer, and (d) proposed method. The notations are the same as Fig.3.

into the neighbored class after being attacked. Fig.6 shows that the robustness of classifier has been greatly improved through defense training, as the quantity of misclassified

adversarial examples (red triangles) are obviously reduced in Fig.6(b)-(d). Note that a robust classifier is supposed to not only alleviate misclassification but draw the adversarial



**FIGURE 8. Robustness of different layers for attention-based feature matching.** The conv4\_3 layer achieves slightly better adversarial accuracy (see solid line in pink) than the proposed layer, nevertheless, its performance spectacularly degrades on standard classification (see solid line in blue). Note that while others layer can well perform on either standard or adversarial classification task, the proposed penultimate layer is able to yield better performance than average on both scenes.

images from false class back to the true one (cluster of blue dots), as shown in Fig.6(d). Thus, our method is able to better resistant to adversarial perturbations by reject adversarial examples that lie on the edge of the false cluster, which corresponding to delusive images that much difficult to recognize.

**D. DESIGN CHOICES**

1) VARIANTS OF FEATURE MATCHING OPERATIONS

Next, we evaluate the variants of feature matching operations described above by comparing the loss surface around the test data points. As shown in Fig.7, each of the feature matching operations ((b)-(d)) is useful, as the loss gets reduced and the surface becomes smoother when compared with the standard model. This suggests that feature matching is a sensible design principle for defending against adversarial attacks. It is noticeable that the proposed method yields much more flatter landscapes and lower losses, which empirically demonstrates the potential of matching the class-specific activation for improving the adversarial robustness.

2) CHOICE OF LAYER

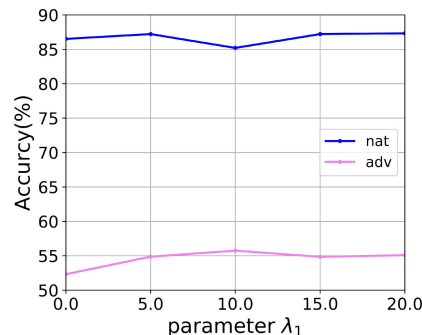
Quantitative experiments have been conducted to validate our choice of layer. We set the preferred layer for feature matching by the experiment on how the adversarial accuracy changes as the layer varies. Results reported in Fig.8 and Table 3 demonstrate that the penultimate layer is effective for attention-based feature matching, while other convolutional layers are less than ideal.

**TABLE 3. Robustness of different layers for attention-based feature matching.**

Layer	conv2_5	conv3_3	conv3_5	conv4_3	Proposed
Adversarial accuracy(%)	54.06	54.31	53.72	60.70	55.74
Standard accuracy(%)	85.98	81.07	85.86	63.87	85.21

3) CHOICE OF PARAMETER VALUES

Parameter  $\lambda_1$  is the weight of the proposed regularizer, serving as a controller on the strength of feature matching operation when applying to adversarial training. Notable gains



**FIGURE 9. Classification accuracy on natural images and adversarial examples under different settings of parameter  $\lambda_1$ .** Setting  $\lambda_1 = 0$  corresponds to adversarial training without our regularizer, while different positive values reflect varied strength of regularization. We recommend  $\lambda_1 = 10$  for a well balance of standard classification performance and adversarial accuracy.

of accuracy on adversarial examples can be achieved by simply setting the parameter  $\lambda_1$  positive, as shown in Fig.9, suggesting that the proposed regularizer is essential to an enhanced defense. We also observed that lower accuracy on clean images accompanies higher adversarial accuracy. Reason is that a provable trade-off exists between standard training and adversarial training [38]. Therefore, we suggest that setting  $\lambda_1 = 10$  for a well balance of standard classification performance and adversarial accuracy.

**V. CONCLUSION AND FUTURE WORK**

In this work, we first perform an empirical analysis to better understand the adversarial attack. It suggests that, whereas adversarial perturbations are restricted to be subtle at pixel level, alternations of the representation at high-level are not hidden. Motivated by this observation, we then present a regularizer that leverages the class activation mapping technique for featuring matching. Extensive empirical experiments show that model trained by the proposed regularization term distinguish itself from the baseline methods on the benchmark datasets CIFAR-10 and CIFAR-100. We also conduct comparative experiments on alternative feature matching operations. The outperformance of the proposed method over other variants suggests that suppression of attention shift enables effective defense against the adversarial examples. Overall, the inspiration drawn from human attention on the adversarial defense has desirably advanced the state-of-the-art techniques. This provides us a broad view on future work that the integration of bio-inspired mechanism into artificial intelligence methodologies has great potential to yield further improvements.

One of the limitations to our work is that the proposed method has not been scaled to and verified on the more challenging benchmark datasets (e.g., ImageNet) or real-world scenario, due to the prohibitively high computational cost of adversarial training. With more available hardware in future we might get closer to the true robustness. Note that the technique we use for generating the attention map is applicable to any CNN-based classifiers, and the attack we employ for



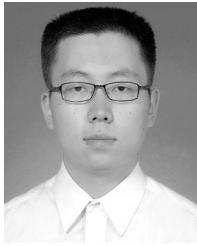
adversarial training is universal. In other words, our method is independent of network architecture and attacks. Thereby, it is conceivable that our approach could be used in conjunction with sufficiently expressive networks and stronger attacks in future to develop even better defense.

## REFERENCES

- [1] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, Oct. 2019.
- [2] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Berlin, Germany: Springer, 2013, pp. 387–402.
- [3] K. Pei, Y. Cao, J. Yang, and S. Jana, "DeepXplore: Automated whitebox testing of deep learning systems," in *Proc. 26th Symp. Operating Syst. Princ.*, Oct. 2017, pp. 1–18.
- [4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [5] H. Wang, C. Mao, H. He, M. Zhao, T. S. Jaakkola, and D. Katabi, "Bidirectional inference networks: A class of deep Bayesian networks for health profiling," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 766–773.
- [6] Y. Tian, K. Pei, S. Jana, and B. Ray, "DeepTest: Automated testing of deep-neural-network-driven autonomous cars," in *Proc. 40th Int. Conf. Softw. Eng.*, May 2018, pp. 303–314.
- [7] S. Wang, K. Pei, J. Whitehouse, J. Yang, and S. Jana, "Formal security analysis of neural networks using symbolic intervals," in *Proc. 27th Secur. Symp. (USENIX Security)*, 2018, pp. 1599–1614.
- [8] C. Mao, K. Lin, T. Yu, and Y. Shen, "A probabilistic learning approach to UWB ranging error mitigation," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.
- [9] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*. [Online]. Available: <http://arxiv.org/abs/1312.6199>
- [10] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*. [Online]. Available: <http://arxiv.org/abs/1412.6572>
- [11] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," 2017, *arXiv:1705.07204*. [Online]. Available: <http://arxiv.org/abs/1705.07204>
- [12] Z. Yan, Y. Guo, and C. Zhang, "Deep defense: Training dnns with improved adversarial robustness," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 419–428.
- [13] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier, "Parseval networks: Improving robustness to adversarial examples," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 854–863.
- [14] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018.
- [15] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
- [16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 618–626.
- [17] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?" *Comput. Vis. Image Understand.*, vol. 163, pp. 90–100, Oct. 2017.
- [18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*. [Online]. Available: <http://arxiv.org/abs/1706.06083>
- [19] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," 2016, *arXiv:1611.01236*. [Online]. Available: <http://arxiv.org/abs/1611.01236>
- [20] D. Meng and H. Chen, "MagNet: A two-pronged defense against adversarial examples," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2017, pp. 135–147.
- [21] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry, "Adversarially robust generalization requires more data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 5014–5026.
- [22] H. Zhang and J. Wang, "Defense against adversarial attacks using feature scattering-based adversarial training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1829–1839.
- [23] H. Kannan, A. Kurakin, and I. Goodfellow, "Adversarial logit pairing," 2018, *arXiv:1803.06373*. [Online]. Available: <http://arxiv.org/abs/1803.06373>
- [24] L. Engstrom, A. Ilyas, and A. Athalye, "Evaluating and understanding the robustness of adversarial logit pairing," 2018, *arXiv:1807.10272*. [Online]. Available: <http://arxiv.org/abs/1807.10272>
- [25] C. Xie, Y. Wu, L. V. D. Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 501–509.
- [26] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*. [Online]. Available: <http://arxiv.org/abs/1503.02531>
- [27] T. Pang, K. Xu, C. Du, N. Chen, and J. Zhu, "Improving adversarial robustness via promoting ensemble diversity," 2019, *arXiv:1901.08846*. [Online]. Available: <http://arxiv.org/abs/1901.08846>
- [28] J. Uesato, J.-B. Alayrac, P.-S. Huang, R. Stanforth, A. Fawzi, and P. Kohli, "Are labels required for improving adversarial robustness?" 2019, *arXiv:1905.13725*. [Online]. Available: <http://arxiv.org/abs/1905.13725>
- [29] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
- [30] A. Mahendran and A. Vedaldi, "Visualizing deep convolutional neural networks using natural pre-images," *Int. J. Comput. Vis.*, vol. 120, no. 3, pp. 233–255, May 2016.
- [31] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 694–711.
- [32] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.
- [33] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, USA, Tech. Rep. TR-2009, 2009.
- [34] C. Mao, Z. Zhong, J. Yang, C. Vondrick, and B. Ray, "Metric learning for adversarial robustness," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 478–489.
- [35] C. Qin, J. Martens, S. Goyal, D. Krishnan, K. Dvijotham, A. Fawzi, S. De, R. Stanforth, and P. Kohli, "Adversarial robustness through local linearization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 13824–13833.
- [36] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.
- [37] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," 2018, *arXiv:1805.12152*. [Online]. Available: <http://arxiv.org/abs/1805.12152>
- [38] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 125–136.
- [39] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. El Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," 2019, *arXiv:1901.08573*. [Online]. Available: <http://arxiv.org/abs/1901.08573>
- [40] J. Uesato, B. O'Donoghue, A. van den Oord, and P. Kohli, "Adversarial risk and the dangers of evaluating against weak attacks," 2018, *arXiv:1802.05666*. [Online]. Available: <http://arxiv.org/abs/1802.05666>



**ZHUORONG LI** received the Ph.D. degree in control science and engineering from the Zhejiang University of Technology. She is currently a Lecturer with the Zhejiang University City College. Her research interests include computational vision, deep learning and corresponding applications, and machine learning.



**CHAO FENG** received the B.E. degree from Shanghai Jiaotong University. He is currently pursuing the M.S. degree with Zhejiang University. His research interests include computational vision, and deep learning and corresponding applications.



**MINGHUI WU** received the Ph.D. degree in computer science from Zhejiang University, China, in 2011. He is currently a Professor with the Zhejiang University City College. He has academic articles published in reputable journals and conferences, including AAAI, KDD, and WWW. His research interests include mobile application and artificial intelligence. He has been rewarded Google Faculty Award three times.



**JIANWEI ZHENG** received the B.Sc. degree, in 2005, and the Ph.D. degree in control theory and control engineering from the Zhejiang University of Technology, China, in 2010. He is currently an Associate Professor with the School of Computer Science and Engineering, Zhejiang University of Technology. He has published more than 60 academic articles in reputable journals and conferences, including the *IEEE TRANSACTIONS ON IMAGE PROCESSING*, *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*, *IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS*, *Neurocomputing*, *Visual Computer*, *Applied Intelligence*, *PCM*, and *CGI*.



**HONGCHUAN YU** (Member, IEEE) received the Ph.D. degree in computer vision from the Institute of Intelligent Machine, Chinese Academy of Sciences, in 2000. He is currently a Principal Academic of computer graphics with the National Centre for Computer Animation, Bournemouth University. His specialties include geometry, graphics, and image processing. He is a Fellow of the High Education of Academy, U.K.

• • •