ROBUST REPRESENTATION LEARNING VIA ASYMMETRIC NEGATIVE CONTRAST AND REVERSE ATTENTION

Anonymous authors
Paper under double-blind review

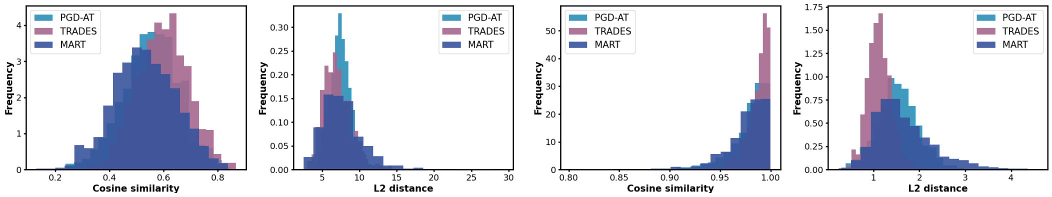
ABSTRACT

Deep neural networks are vulnerable to adversarial noise. Adversarial training (AT) has been demonstrated to be the most effective defense strategy to protect neural networks from being fooled. However, we find AT omits to learning robust features, resulting in poor performance of adversarial robustness. To address this issue, we highlight two characteristics of robust representation: *(1) exclusion: the feature of natural examples keeps away from that of other classes; (2) alignment: the feature of natural and corresponding adversarial examples is close to each other.* These motivate us to propose a generic framework of AT to gain robust representation, by the asymmetric negative contrast and reverse attention. Specifically, we design an asymmetric negative contrast based on predicted probabilities, to push away examples of different classes in the feature space. Moreover, we propose to weight feature by parameters of the linear classifier as the reverse attention, to obtain class-aware feature and pull close the feature of the same class. Empirical evaluations on three benchmark datasets show our methods greatly advance the robustness of AT and achieve state-of-the-art performance.

1 INTRODUCTION

Deep neural networks (DNNs) have achieved great success in academia and industry, but they are easily fooled by carefully crafted adversarial examples to output incorrect results (Goodfellow et al., 2014), which leads to potential threats and insecurity in the application. Given a naturally trained DNN and a natural example, an adversarial example can be generated by adding small perturbations to the natural example. Adversarial examples can always fool models to make incorrect output. At the same time, it is difficult to distinguish adversarial examples from natural examples by human eyes. In recent years, there are many researches exploring the generation of adversarial examples to cheat models in various fields, including image classification (Goodfellow et al., 2014; Madry et al., 2017; Carlini & Wagner, 2016; Croce & Hein, 2020), object detection (Xie et al., 2017; Chen et al., 2021b), natural language processing (Morris et al., 2020; Boucher et al., 2022), semantic segmentation (Nesti et al., 2022; Luo et al., 2022), etc. The vulnerability of DNNs has aroused common concerns on adversarial robustness.

Many empirical defense methods have been proposed to protect DNNs from adversarial perturbations, such as adversarial training (AT) (Madry et al., 2017; Zhang et al., 2019; Wang et al., 2020; Huang et al., 2020; Zhou et al., 2022; Zhang et al., 2020; Wu et al., 2020), image denoising (Liao et al., 2018), defensive distillation (Zhao et al., 2022; Chen et al., 2021a) and so on. The mainstream view is that AT is the most effective defense, which has a training process of a two-sided game. The "attacker" crafts perturbation dynamically to generate adversarial data to cheat the "defender", and the "defender" minimizes the loss function against adversarial samples to improve the robustness of models. Existing work (Zhao et al., 2022; Chen et al., 2021a; Zhang et al., 2020; Dong et al., 2021; Huang et al., 2020; Jin et al., 2022; Zhou et al., 2022) has improved the effectiveness of AT in many aspects, but few studies pay attention to learning robust feature. The overlook may lead to potential threats in the feature space of AT models, which harms robust classification. Besides, there are no criteria for robust feature. In addition, adversarial contrastive learning (ACL) and robust feature selection (RFS) are techniques to optimize feature distribution. ACL (Kim et al., 2020; Fan et al., 2021; Yu et al., 2022) is a kind of contrast learning (CL) (Chen et al., 2020; He et al., 2020; Grill et al., 2020) that extends to AT. RFS mostly modifies the architecture of models (Xiao et al.,



(a) cosine similarity of natural examples and OEs

(b) L2 distance of natural examples and OEs

(c) cosine similarity of natural examples and AEs

(d) L2 distance of natural examples and AEs

Figure 1: Frequency histograms of the L_2 distance and cosine similarity of the feature that belongs to natural examples, AEs and OEs. We train ResNet-18 models on CIFAR-10 with three AT methods: PDG-AT (Madry et al., 2017), TRADES (Zhang et al., 2019) and MART (Wang et al., 2020). We use all samples labeled as class 0 in the test set as natural examples and generate AEs by PGD-10.

2019; Bai et al., 2021; Yan et al., 2021) to select important feature. However, the target problems of them are not to learn robust feature.

To demonstrate AT is indeed deficient in the representation which causes limited adversarial robustness, we conduct a simple experiment. We choose the L_2 distance and cosine similarity as metrics. And we measure the distance and similarity of the feature between natural examples, adversarial examples **(AEs)** and examples of other classes **(OEs)**. The frequency histograms of the distance and similarity are shown in Figure 1. Figure 1 (a) and Figure 1 (b) show that the cosine similarity of the feature between natural examples and OEs shows a bell-shaped distribution between 0.4 and 0.8, and the L_2 distance shows a skewed distribution between 2.0 and 12.0, which indicates there are very close pairs of natural examples and OEs that are not distinguished in the feature space. In Figure 1 (c) and Figure 1 (d), it is shown that there is a skewed distribution between 0.9 and 0.99 for the cosine similarity of the feature between natural examples and AEs, and a skewed distribution between 0.5 and 2.5 for the L_2 distance, which indicates that the feature of natural examples and AEs is not adequately aligned. Thus, there is still large room for optimization of the feature of AT.

Based on the observation, we propose two characteristics of robust feature: *exclusion: the feature of natural examples keeps away from that of other classes; alignment: the feature of natural and corresponding adversarial samples is close to each other*. First, *exclusion* confirms the separability between different classes and avoids confusion in the feature space, which makes it hard to fool the model because the feature of different classes keeps a large distance. Second, *alignment* ensures the feature of natural examples is aligned with adversarial one, which guarantees the predicted results of the natural and adversarial examples of the same instances are also highly consistent. And it helps to narrow the gap between robust accuracy and clean accuracy.

To address the issue, we propose an AT framework to concentrate on robust representation with the guidance of the two characteristics. Specifically, we suggest two strategies to meet the characteristics, respectively. For *exclusion*, we propose an asymmetric negative contrast based on predicted probabilities, which freezes natural examples and pushes away OEs by reducing the confidence of the predicted class when predicted classes of natural examples and OEs are consistent. For *alignment*, we use the reverse attention to weight feature by parameters of the linear classifier corresponding to target classes, which contains the importance of feature to target classes during classification. Because the feature of the same class gets the same weighting and feature of different classes is weighted disparately, natural examples and AEs become close to each other in the feature space. Empirical evaluations show that AT methods combined with our framework can greatly enhance robustness, which means the neglect of learning robust feature is one of the main reasons for the poor robust performance of AT. In a word, we propose a generic AT framework with the Asymmetric Negative Contrast and Reverse Attention (**ANCRA**), to learn robust representation and advance robustness. Our main contributions are summarized as follows:

- We suggest improving adversarial training from the perspective of learning robust feature, and two characteristics are highlighted as criteria of optimizing robust representation.
- We propose a generic framework of adversarial training, termed as ANCRA, to obtain robust feature by the asymmetric negative contrast and reverse attention, with the guidance of two characteristics of robust feature. It can be easily combined with other defense methods.

- Empirical evaluations show our framework can obtain robust feature and greatly improve adversarial robustness, which achieves state-of-the-art performances on CIFAR-10, CIFAR-100 and Tiny-ImageNet.

2 Related work

Adversarial training Madry et al. (2017) propose PGD attack and PGD-based adversarial training, forcing the model to correctly classify adversarial samples within the epsilon sphere during training to obtain robustness, which is the pioneer of adversarial learning. Zhang et al. (2019) propose to learn both natural and adversarial samples and reduce the divergence of classification distribution of both to reduce the difference between robust accuracy and natural accuracy. Wang et al. (2020) find that misclassified samples during training harm robustness significantly, and propose to improve the model's attention to misclassification by adaptive weights. Zhang et al. (2020) propose to replace fixed attack steps with attack steps that just cross the decision boundary, and improve the natural accuracy by appropriately reducing the number of attack iterations. Huang et al. (2020) replace labels with soft labels predicted by the model and adaptively reduce the weight of misclassification loss to alleviate robust overfitting problem. Dong et al. (2021) also propose a similar idea of softening labels and explain the different effects of hard and soft labels on robustness by investigating the memory behavior of the model for random noisy labels. Chen et al. (2021a) propose random weight smoothing and self-training based on knowledge distillation, which greatly improves the natural and robust accuracy. Zhou et al. (2022) embed a label transition matrix into models to infer natural labels from adversarial noise. However, little work has been done to improve AT from the perspective of robust feature learning. Our work shows AT indeed has defects in the feature distribution, and strategies proposed to learn robust feature can greatly advance robustness, which indicates the neglect of robust representation results in poor robust performance of AT.

Adversarial contrastive learning Kim et al. (2020) propose to maximize and minimize the contrastive loss for training. Fan et al. (2021) notice that the robustness of ACL relies on fine-tuning, and pseudo labels and high-frequency information can advance robustness. Kucer et al. find that the direct combination of self-supervised learning and AT penalizes non-robust accuracy. Bui et al. (2021) propose some strategies to select positive and negative examples based on predicted classes and labels. Yu et al. (2022) find the instance-level identity confusion problem brought by positive contrast and address it by asymmetric methods. These methods motivate us to further consider how to obtain robust feature by contrast mechanism. We design a new negative contrast to push away natural and negative examples and mitigate the confusion caused by negative contrast.

Robust feature selection Xiao et al. (2019) take the maximum k feature values in each activation layer to increase adversarial robustness. Zoran et al. (2020) use a spatial attention mechanism to identify important regions of the feature map. Bai et al. (2021) propose to suppress redundant feature channels and dynamically activate feature channels with the parameters of additional components. Yan et al. (2021) propose to amplify the top-k activated feature channels. Existing work has shown enlarging important feature channels is beneficial for robustness, but most approaches rely on extra model components and do not explain the reason. We propose the reverse attention to weight feature by class information without any extra components, and explain it by *alignment* of feature.

3 Methodology

This section explains the instantiation of our AT framework from the perspective of the two characteristics of robust feature. To meet *exclusion*, we design an asymmetric negative contrast based on predicted probabilities to push away the feature of natural examples and OEs. To confirm *alignment*, we propose the reverse attention to weight the feature of the same class, by the corresponding weight of the targeted class in parameters of the linear classifier, so that the feature of natural examples and AEs is aligned and the gap of the feature between natural examples and AEs becomes small.

3.1 Notations

In this paper, capital letters indicate random variables or vectors, while lowercase letters represent their realizations. We define the function for classification as $f(\cdot)$. It can be parameterized by DNNs.

$Linear(\cdot)$ is the linear classifier with a weight of Ω (C, R), in which C denotes the class number and R denotes the channel number of the feature map. $g(\cdot)$ is the feature extractor, i.e., the rest model without $Linear(\cdot)$. Let $\mathcal{B} = \{x_i, y_i\}_i^N$ be a batch of natural samples where x_i is labeled by y_i. x^a denotes adversarial examples **(AEs)**, x^o denotes negative examples randomly selected from other classes **(OEs)**. Given an adversarial transformation $\mathcal{T}_a$ from an adversary $\mathcal{A}$ (e.g., PGD attack (Madry et al., 2017)), and a strategy $\mathcal{T}_o$ for selection or generation of OEs. For data, we consider a positive pair PP=$\{x_i, x_i^a | x_i \in \mathcal{B}, x_i^a = \mathcal{T}_a(x_i)\}_i^N$. We define a negative pair NP=$\{x_i, x_i^o | x_i \in \mathcal{B}, x_i^o = \mathcal{T}_o(x_i)\}_i^N$. Let $\mathbb{N}(x, \epsilon)$ represents the neighborhood of $x : \{\tilde{x} : \|\tilde{x} - x\| \leq \epsilon\}$, where ϵ is the perturbation budget. For an input x_i, we consider its feature z_i before $Linear(\cdot)$, the probability vector $p_i = softmax(f(x_i))$ and predicted class $h_i = argmax(p_i)$, respectively.

3.2 Adversarial training with asymmetric negative contrast

First, we promote AT to learn robust representation that meets *exclusion*. Notice that ACL has the contrastive loss (van den Oord et al., 2018) to maximize the consistency of PPs and to minimize the consistency of NPs. Motivated by the contrast mechanism, we consider designing a new negative contrast and combining it with AT loss, which creates a repulsive action between NPs when minimizing the whole loss. Thus, we propose a generic pattern of AT loss with a negative contrast.

$$\mathcal{L}^{\mathrm{CAL}}(x, y, x^a, x^o) = \mathcal{L}^{\mathrm{AT}}(x, x^a, y) + \zeta \cdot \mathrm{Sim}\left(x, x^o\right), \tag{1}$$

Where x denotes natural examples with labels y, x^a denotes AEs, x^o denotes OEs, Sim is a similarity function, and ζ is the weight of Sim. Generated by maximizing $\mathcal{L}_{CE}$, AEs typically have wrong predicted classes. The generation is as follows:

$$x_{t+1}^a := \prod_{\mathbb{N}(x,\epsilon)} \left(x_t^a + \epsilon \,\mathrm{sign}\left(\nabla_x \mathcal{L}_{CE}\left(\left(f\left(x_t^a\right), y\right)\right)\right), \tag{2}$$

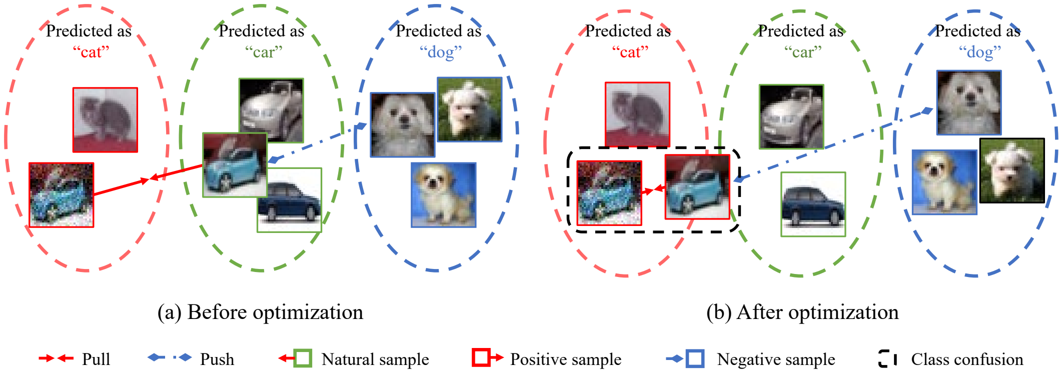
where ϵ denotes the L_∞-norm of perturbation budget, x_t^a denotes adversarial samples after the tth attack iteration, $\mathbf{\Pi}$ denotes a clamp function, $sign$ denotes a sign function, $\mathcal{L}_{CE}$ denotes the cross-entropy loss and $\nabla_x \mathcal{L}_{CE}$ denotes the gradient of $\mathcal{L}_{CE}$ with respect to x. When minimizing Equation 1, $\mathcal{L}^{\mathrm{AT}}$ learns to classify natural examples and AEs correctly, and additional negative contrast promotes the inconsistency of NPs, which keeps the feature of NPs away from each other to ensure *exclusion*. We will further discuss the function and problem of the negative contrast.

(Yu et al., 2022) have indicated that when the predicted classes of the adversarial positive examples (i.e., AEs) and negative samples (i.e., OEs) are the same, there is a conflict led by the positive contrast, resulting in wrong classification. On this basis, we find a similar conflict can also be caused by the negative contrast when their predicted classes are different, which is called class confusion. We show a practical instance in Figure 2. When optimizing the class space, the positive example pulls its natural example close to the wrong class. The negative example pushes the natural example to leave the initial class. With these actions, the training process suffers from class confusion, leading to natural examples moving toward the wrong class space, which does harm to *exclusion*.

To alleviate the problem of class confusion, We should reasonably control the repulsion of negative contrast. We propose an asymmetric method of the negative contrast, $\mathrm{Sim}^\alpha(x, x^o)$, to decouple the repulsive force into two parts. It contains a one-side push from the natural example to the OE and a one-side push from the OE to the natural example, given by:

$$\mathrm{Sim}^\alpha(x, x^o) = \alpha \cdot \overline{\mathrm{Sim}}(x, x^o) + (1 - \alpha) \cdot \overline{\mathrm{Sim}}(x^o, x), \tag{3}$$

where $\overline{\mathrm{Sim}}(x, x^o)$ denotes the one-sided similarity of x and x^o. When minimizing $\overline{\mathrm{Sim}}(x, x^o)$, we stop the back-propagation gradient of x and only move x^o away from x. α denotes the weighting factor to adjust the magnitude of the two repulsive forces. When α = 0, negative samples are frozen and only the feature of natural samples is optimized to be pushed far away from the feature of negative



(a) Before optimization

(b) After optimization

Figure 2: Illustrations of class confusion. In each circle, data points have the same predicted class. In (a), positive examples are adversarial examples and located in the wrong predicted class different from natural examples. The positive contrast pulls natural examples to move toward the wrong class, and the negative contrast pushes natural examples to leave the initial class. Finally, natural examples come to the decision boundary and even into the wrong class easily as (b) shows.

samples. As α increases, the natural sample becomes more repulsive to the negative sample and the negative sample pushes the natural example less. To mitigate the class confusion problem, we should choose α that tends to 1 to reduce the repulsive force from the negative sample to the natural example, to prevent the natural example from being pushed into the wrong class. Experiments show that α =1 leads to the best performance (provided in Appendix A.2), which optimizes NPs by only pushing off negative samples and follows what we have expected.

Then we propose the negative contrast based on predicted probabilities, $\mathrm{Sim}^{\alpha}_{cc}(x, x^o)$, to measure the repulsive force of NPs pushing away from each other. It pushes away NPs by decreasing the corresponding probabilities of the predicted classes when the predicted classes of NPs are consistent.

$$\mathrm{Sim}^{\alpha}_{cc}(x, x^o) = \frac{1}{\|\mathcal{B}_i\|} \sum_{i=1}^{n} \mathbb{I}\left(h_i = h_i^o\right) \cdot \left[\alpha \sqrt{\hat{p}_i(h_i) \cdot p_i^o(h_i)} + (1-\alpha) \sqrt{p_i(h_i) \cdot \hat{p}_i^o(h_i)}\right], \tag{4}$$

where $\|\mathcal{B}_i\|$ denotes the batch size, $\mathbb{I}(\cdot)$ denotes the Indicator function and $\hat{p}$ denotes freezing the back-propagation gradient of p. h_i and h_i^n denote the predicted classes of the NP. And p_i and p_i^n denote the probability vectors of the NP. Under the negative contrast, the model pushes the natural example in the direction away from the predicted class of the OE and pushes the OE in the direction away from the predicted class of the natural example when and only when two predicted classes of the NP are consistent. This ensures that the action of *exclusion* not only pushes away the feature of NPs in the feature space, but also reduces the probabilities of NPs in the incorrect class. Since the negative contrast has only directions to reduce the confidence and no explicit directions to increase the confidence, it does not create any actions to push the natural example into the feature space of wrong classes even in the scenario of class confusion, which can effectively alleviate the problem.

3.3 Adversarial training with reverse Attention

Second, we continue to improve AT to learn robust representation that meets *alignment*. Motivated by Bai et al. (2021); Yan et al. (2021), we utilize the values of linear weight to denote the importance of feature channels to targeted classes. We exploit the importance of feature channels to align the examples in the same classes and pull close the feature of PPs, which is named by reverse attention. To be specific, we take the Hadamard product (Kronecker product) of the partial parameters of the classifier Ω^j and the feature vector z. 'partial parameters' means those weights of the linear layer that are used to calculate the probability of the target class. Because reverse attention weights the feature of PPs by the same parameters, it helps *alignment*. Given by:

$$z'_i = \begin{cases} z_i \odot \omega^{i,y}, & \text{(training phase)} \\ z_i \odot \omega^{i,h(x)}, & \text{(testing phase)} \end{cases} \tag{5}$$

Where z denotes the feature vector, z_i denotes the importance of the ith feature channel, $\omega^{i,j}$ denotes the linear parameters of the ith feature channel to the jth class, $\odot$ denotes the Hadamard product operation, which is the method of multiplying two matrices of the same size element by element.

As shown in Figure3, We add the reverse attention to the last layer of the model and then get auxiliary probability vectors as many as the blocks in the layer. In the ith block, the unweighted feature vector z^i goes through $Linear(\cdot)$ and the auxiliary probability vector p^i outputs, unless we get the final probability vector p'. Suppose there are n blocks in the last layer, we will use p^{n-2}, p^{n-1}, p' to calculate losses and add them together. During the training phase, we use the true label y as an indicator to determine the importance of channels. In the testing phase, since the true label is not available, we simply choose a sub-vector of the linear weight by the predicted class $h(x)$ as the importance of channels. The model with the reverse attention does not need any extra modules, but module interactions are changed.

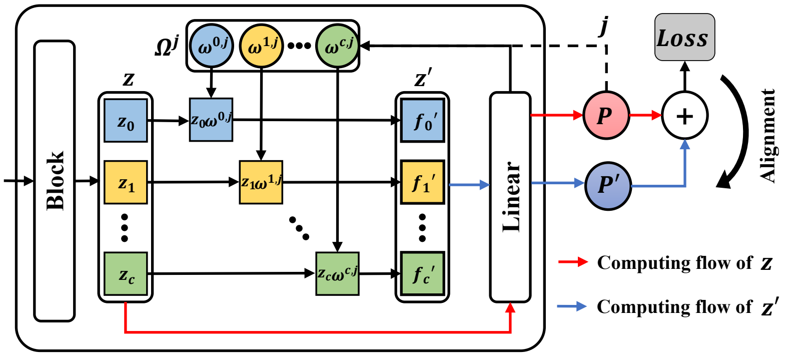


Figure 3: An illustration of reverse attention (RA). z and z' denote the feature vector before and after weighting, Ω^j denotes the linear parameters of the target class j. p and p' denote the probability vector before and after weighting. In the calculation, z is multiplied by Ω^j to get z'. And we calculate p' and p and add the losses of them to get the total loss. When minimizing the loss, the similarity of examples in the same class becomes bigger, causing attractive forces to each other.

Let's make a detailed analysis and explanation of the principle of this method. In the model, the feature extractor captures the representation that contains enough information to classify, and the linear layer establishes a relationship from feature to predicted classes. The probability of the predicted class equals the sum of the product of linear weight corresponding to predicted class and feature vector. In this premise, the linear layer learns to correctly increase the probability of the label class and decrease other probabilities when training. Thus it can gradually recognize which feature channels are important for specific classes, and keep large weight values for those feature channels. On this basis, we propose reverse attention to utilize its parameters containing feature importance to improve feature. The feature vectors are multiplied by the parameters of the target class, which can change the magnitude of each feature channel adaptively according to the feature importance, acting as attention with the guidance of the linear layer. From the perspective of the feature itself, the important channels in the feature vector are boosted and the redundant channels are weakened after the attention. Therefore, the feature value contributing to the target class will become larger, which is helpful for correct classification. From the perspective of the overall feature distribution, reverse attention can induce beneficial changes in the feature distribution. Since the linear layer is unique in the model, different examples in the same class share the same linear weights. Feature vectors with the same target class(e.g., examples in PPs) get the same weighting and become more similar. Moreover, feature vectors with different target classes(e.g., examples in NPs) are weighted by different parameters, and the weighted feature distributions may become more inconsistent. Therefore, the reverse attention guides the alignment of the feature of the examples in the same class, pulling the feature of PPs closer and pushing the feature of NPs far away, which benefits *alignment* and drops by to promote *exclusion*. The aligned feature has similar activations in every feature channel, which helps the model narrow the gap between feature of natural examples and AEs.

4 Experiments

In order to demonstrate the effectiveness of the proposed approach, we show feature distribution of trained models first. Then we evaluate our framework against white-box attacks and adaptive attacks, and make a comparison with other defense methods. We conduct experiments across different datasets and models. Because our methods are compatible with existing AT techniques and can be easily incorporated in a plug-and-play manner, we choose three baselines to combine with our framework for evaluation: PGD-AT-ANCRA (Madry et al., 2017), TRADES-ANCRA (Zhang et al., 2019), and MART-ANCRA (Wang et al., 2020).

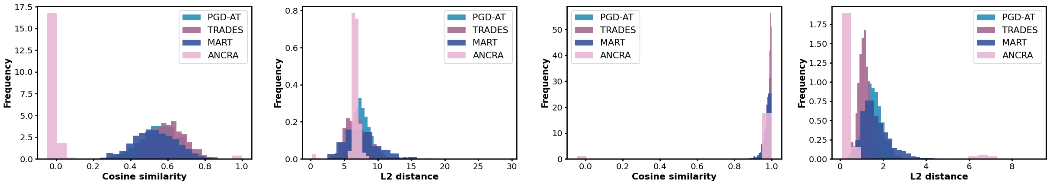
4.1 Settings

Implementation On CIFAR-10 and CIFAR-100 (Krizhevsky, 2009), we train ResNet-18 (He et al., 2016a) with a weight decay of 2.0×10^{-4}. We adopt the SGD optimizer with a learning rate of 0.01, epochs of 120 and a batch size of 128 as Wang et al. (2020). For the trade-off hyperparameters β, we use 6.0 in TRADES[1] and 5.0 in MART, following the original setting in their papers. For other hyperparameters, we tune the values based on TRADES-ANCRA. We generate AEs for training by L_∞-norm PGD (Madry et al., 2017), with a step size of 0.007, an attack iterations of 10 and a perturbation budget of 8/255. We use a single NVIDIA A100 and two GTX 2080 Ti.

Baseline We compare the proposed PGD-AT-ANCRA, TRADES-ANCRA, and MART-ANCRA with the popular baselines: PGD-AT (Madry et al., 2017), TRADES (Zhang et al., 2019), MART (Wang et al., 2020) and SAT (Huang et al., 2020). Moreover, we also choose three state-of-the-art methods: AWP (Wu et al., 2020), S2O (Jin et al., 2022) and UDR (Bui et al., 2022). We keep the same settings among all the baselines with our settings and follow their original hyperparameters.

Evaluation We choose several adversarial attacks to attack the target models, including PGD (Madry et al., 2017), FGSM (Goodfellow et al., 2014), C&W (Carlini & Wagner, 2016) and AutoAttack (Croce & Hein, 2020) which is a powerful and reliable attack and an ensemble attack with three white-box attacks and one black-box attack. We notice that our methods use the auxiliary probability vector p in the training and testing phase, so we design two scenarios: 1) train with p and test without p; 2) train with p and test with p. 1) denotes evaluation against white-box attacks and 2) denotes evaluation against adaptive attacks. Following the default setting of AT, the max perturbation strength is set as 8. / 255. for all attack methods under the L_∞. The attack iterations of PGD and C&W are 40 (i.e., PGD-40), and the step size of FGSM is 8. / 255. unlike 0.007 for other attacks. The clean accuracy and robust accuracy are used as the evaluation metrics.

4.2 Comparison results of feature distribution



(a) Cosine similarity of NP's feature **(b) L2 distance of NP's feature** **(c) Cosine similarity of PP's feature** **(d) L2 distance of PP's feature**

Figure 4: Frequency histograms of the L_2 distance and cosine similarity of feature of natural examples, AEs and OEs. We train ResNet-18 models on CIFAR-10 with four defense techniques: PDG-AT, TRADES, MART and TRADES-ANCRA. Other details are the same with Figure 1

Frequency histograms of feature distribution are shown in Figure 4. It is shown that our methods can greatly improve feature distribution, which follows the characteristics of *exclusion* and *alignment*. In Figure 4 (a) and Figure 4 (b), it shows that the cosine similarity of the model trained by our method between natural examples and OEs shows a skewed distribution between -0.05 and 0.1, and the L_2 distance with our method shows a bell-shaped distribution between 5.5 and 10.0, which indicates natural examples and OEs have been fully distinguished in the feature space and *exclusion*

[1] Unlike vanilla TRADES, we maximize the CE loss to generate adversarial examples as PGD.

has been met. In Figure 4 (c) and Figure 4 (d), it shows that in the model trained by our method, there is a uniform distribution between 0.95 and 0.99 for the cosine similarity of the feature between natural examples and AEs, and a skewed distribution between 0.05 and 1.5 for the L_2 distance of the feature, which indicates the feature between natural examples and AEs is very close to each other and *alignment* has been confirmed. Thus, our framework successfully helps AT to obtain robust feature. More feature visualizations are provided in Appendix A.3.

4.3 Comparison results against white-box attacks

Table 1: Robustness (%) against white-box attacks. Nat denotes clean accuracy. PGD denotes robust accuracy against PGD-40. FGSM denotes robust accuracy against FGSM. C&W denotes robust accuracy against C&W. AA denotes robust accuracy against AutoAttack. Mean denotes average robust accuracy. The variation of accuracy $\leq 1.7\%$. We show the most successful defense with **bold**.

Defense	CIFAR-10						CIFAR-100					
	Nat	PGD	FGSM	C&W	AA	Mean	Nat	PGD	FGSM	C&W	AA	Mean
PGD-AT	80.90	44.35	58.41	46.72	42.14	47.91	56.21	19.41	30.00	41.76	17.76	27.23
TRADES	78.92	48.40	59.60	47.59	45.44	50.26	53.46	25.37	32.97	43.59	21.35	30.82
MART	79.03	48.90	60.86	45.92	43.88	49.89	53.26	25.06	33.35	38.07	21.04	29.38
SAT	63.28	43.57	50.13	47.47	39.72	45.22	42.55	23.30	28.36	41.03	18.73	27.86
AWP	76.38	48.88	57.47	48.22	44.65	49.81	54.53	27.35	34.47	44.91	21.98	31.18
S2O	40.09	24.05	29.76	47.00	44.00	36.20	26.66	13.11	16.83	43.00	21.00	23.49
UDR	57.80	39.79	45.02	46.92	34.73	41.62	33.63	20.61	24.19	33.77	16.41	23.75
PGD-AT-ANCRA	**85.10**	**89.03**	87.00	**89.23**	59.15	81.10	59.73	58.10	58.45	58.58	34.44	52.39
TRADES-ANCRA	81.70	**82.96**	82.74	83.01	**59.70**	77.10	53.73	51.24	52.17	52.55	**35.81**	47.94
MART-ANCRA	84.88	88.56	**87.95**	88.77	59.62	**81.23**	**60.10**	**58.40**	**58.74**	**59.41**	35.05	**52.90**

We have conducted experiments on ResNet-18 to evaluate different defenses under white-box attacks. The results are shown in Table 1. First, on CIFAR-10, our approaches improve the clean accuracies of based approaches by 5.2%, 3.2% and 5.9%, and also improve the robust performance under all the attacks (e.g., increase by 44.7%, 34.6% and 39.7% against PGD). Compared with state-of-the-art defenses, the robust accuracies of our methods are almost two times as large as theirs (e.g., 81.23% > 49.81%). Second, on CIFAR-100, our approaches also greatly improve the robustness and advance the clean accuracies. The clean accuracies of our methods have been increased by 3.5%, 0.3% and 6.8% compared with based methods, and the lowest average robust accuracy of ours is larger than the best one among other methods by 16.8%. We also train PreActResNet-18 (He et al., 2016b) models on Tiny-ImageNet (Deng et al., 2009). As shown in Table 2, our methods made obvious progress in robustness and generation compared with baselines. In general, our three approaches gain the best performance both in the natural and attacked scenarios. To our surprise, MART-ANCRA and PGD-ANCRA rather than TRADES-ANCRA gain the best performance in a lot of cases without hyper-parameter tuning. More results are provided in Appendix A.5, A.7.

Table 2: Clean and robust accuracy (%) of PreActResNet-18 on Tiny-ImageNet.

Defense	PGD-AT	PGD-AT-ANCRA	TRADES	TRADES-ANCRA	MART	MART-ANCRA
Nat	41.31	43.02	37.27	38.94	38.94	43.83
PGD	10.28	29.79	16.30	31.24	14.78	31.44

4.4 Comparison results against adaptive attacks

We report the performance on CIFAR-10 against adaptive attacks with p^i to evaluate the robustness. Besides, we report vanilla TRADES as a baseline. As shown in Table 3 and Table 4, the robust accuracies of our method against adaptive attacks are larger than those of the baseline against vanilla attacks. e.g., robustness on ResNet-18 against adaptive PGD is higher than the baseline by 13.28% and robustness on WideResNet-34-10 (Zagoruyko & Komodakis, 2016) against adaptive PGD is higher than the baseline by 2.88%. The robustness under adaptive AutoAttack has increased slightly, but not by a significant margin (0.74%, 1.20%). We will discuss the reasons in the Limitation. The results indicate that our approaches can still maintain superb performance under adaptive attacks.

Table 3: Robust accuracy (%) against adaptive attacks of ResNet-18.

Defense	Adaptive Attacks		
	PGD	FGSM	C&W
TRADES	48.40	59.60	47.59
TRADES-ANCRA	61.68	61.56	72.36
PGD-AT-ANCRA	54.43	58.23	66.36
MART-ANCRA	56.96	60.43	71.06

Table 4: Robust accuracy (%) against adaptive attacks of WideResNet (WRN) models.

Defense	Model	Adaptive Attacks	
		PGD	AA
TRADES	WRN-28-10	57.08	51.11
TRADES-ANCRA	WRN-28-10	58.60	51.85
TRADES	WRN-34-10	56.47	50.79
TRADES-ANCRA	WRN-34-10	59.35	51.99

4.5 Ablation studies

We train four models by TRADES, TRADES with the asymmetric negative contrast (TRADES-ANC), TRADES with the reverse attention (TRADES-RA) and TRADES-ANCRA, respectively. As shown in Table 5, when incorporating individual ANC or RA, the performance of robustness and generalization has been improved compared with vanilla TRADES. Besides, when TRADES-ANCRA is compared with other methods, the clean accuracy and robust accuracies against all the attacks except FGSM have been enhanced, which indicates that the two strategies are compatible and the combination can alleviate the side effects of independent methods.

Table 5: Clean and robust accuracy (%) of ResNet-18 trained by TRADES, TRADES-ANC, TRADES-RA and TRADES-ANCRA against various attacks.

Defense	Nat	PGD	FGSM	C&W
TRADES	78.92	48.40	59.60	47.59
TRADES-ANC	80.77	54.18	63.44	49.84
TRADES-RA	80.46	61.59	61.48	72.15
TRADES-ANCRA	81.70	61.68	61.56	72.36

Table 6: Clean and robust accuracy (%) of all the probability vectors trained by TRADES-ANCRA. "Final PV wo RA" means we remove reverse attention and then load trained parameters[2]to test it.

Probability Vector (PV)	Nat	PGD	Adaptive PGD
Auxiliary PV p^0	81.81	83.52	62.25
Auxiliary PV p^1	81.81	83.49	62.23
Final PV p'	81.81	83.47	62.24
Final PV wo RA p''	59.77	58.53	52.81

4.6 Limitation

Because the weights for reverse attention are determined by predicted classes, the wrong predicted classes may lead to the wrong weighted feature and degraded performance. As shown in Table 6, the final predicted results and intermediate predicted labels remain highly consistent. Fortunately, Table 3 has indicated that the high dependence on predicted classes does not significantly affect performance. We will further study this limitation and hope to improve it in the future.

5 Conclusion

This work addresses the overlook of robust representation learning in the adversarial training by a generic AT framework with the asymmetric negative contrast and reverse attention. We propose two characteristics of robust feature to guide the improvement of AT, i.e., *exclusion* and *alignment*. Specifically, the asymmetric negative contrast based on probabilities freezes natural examples, and only pushes away examples of other classes in the feature space. Besides, the reverse attention weights feature by the parameters of the linear classifier, to provide class information and align feature of the same class. Our framework can be used in a plug-and-play manner with other defense methods. Analysis and empirical evaluations demonstrate that our framework can obtain robust feature and greatly improve robustness and generalization.

References

Sravanti Addepalli, Samyak Jain, Gaurang Sriramanan, and R Venkatesh Babu. Scaling adversarial training to large perturbation bounds. In *European Conference on Computer Vision*, pp. 301–316. Springer, 2022a.

[2] the parameters trained by TRADES-ANCRA

Sravanti Addepalli, Samyak Jain, et al. Efficient and effective augmentation strategy for adversarial training. *Advances in Neural Information Processing Systems*, 35:1488–1501, 2022b.

Yang Bai, Yuyuan Zeng, Yong Jiang, Shu-Tao Xia, Xingjun Ma, and Yisen Wang. Improving adversarial robustness via channel-wise activation suppressing. *arXiv preprint arXiv:2103.08307*, 2021.

Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. Bad characters: Imperceptible nlp attacks. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1987–2004, 2022. doi: 10.1109/SP46214.2022.9833641.

Anh Bui, Trung Le, He Zhao, Paul Montague, Seyit Camtepe, and Dinh Phung. Understanding and achieving efficient robustness with adversarial supervised contrastive learning. *arXiv preprint arXiv:2101.10027*, 2021.

Tuan Anh Bui, Trung Le, Quan Hung Tran, He Zhao, and Dinh Q. Phung. A unified wasserstein distributional robustness framework for adversarial training. *CoRR*, abs/2202.13437, 2022.

Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2016.

Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*, 2021a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020.

Xiangning Chen, Cihang Xie, Mingxing Tan, Li Zhang, Cho-Jui Hsieh, and Boqing Gong. Robust and accurate object detection via adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16622–16631, 2021b.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Yinpeng Dong, Ke Xu, Xiao Yang, Tianyu Pang, Zhijie Deng, Hang Su, and Jun Zhu. Exploring memorization in adversarial training. *arXiv preprint arXiv:2106.01606*, 2021.

Lijie Fan, Sijia Liu, Pin-Yu Chen, Gaoyuan Zhang, and Chuang Gan. When does contrastive learning preserve adversarial robustness from pretraining to finetuning? *Advances in neural information processing systems*, 34:21480–21492, 2021.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent a new approach to self-supervised learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016a. doi: 10.1109/CVPR.2016.90.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 630–645, Cham, 2016b. Springer International Publishing. ISBN 978-3-319-46493-0.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020. doi: 10.1109/CVPR42600.2020.00975.

Lang Huang, Chao Zhang, and Hongyang Zhang. Self-adaptive training: beyond empirical risk minimization. *Advances in neural information processing systems*, 33:19365–19376, 2020.

Gaojie Jin, Xinping Yi, Wei Huang, Sven Schewe, and Xiaowei Huang. Enhancing adversarial training with second-order statistics of weights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15273–15283, 2022.

Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:2983–2994, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Michal Kucer, Diane Oyen, and Garrett Kenyon. When does visual self-supervision aid adversarial training in improving adversarial robustness?

Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1778–1787, 2018. doi: 10.1109/CVPR.2018.00191.

Yawei Luo, Ping Liu, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Category-level adversarial adaptation for semantic segmentation using purified features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):3940–3956, 2022. doi: 10.1109/TPAMI.2021.3064379.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Leland McInnes and John Healy. Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv*, abs/1802.03426, 2018.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, October 2020. Association for Computational Linguistics.

Federico Nesti, Giulio Rossolini, Saasha Nair, Alessandro Biondi, and Giorgio Buttazzo. Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2280–2289, 2022.

Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? *arXiv preprint arXiv:2104.09425*, 2021.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020.

Chang Xiao, Peilin Zhong, and Changxi Zheng. Enhancing adversarial defense by k-winners-take-all. *arXiv preprint arXiv:1905.10510*, 2019.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 1369–1378, 2017.

Hanshu Yan, Jingfeng Zhang, Gang Niu, Jiashi Feng, Vincent Tan, and Masashi Sugiyama. Cifs: Improving adversarial robustness of cnns via channel-wise importance-based feature selection. In *International Conference on Machine Learning*, pp. 11693–11703. PMLR, 2021.

Qiying Yu, Jieming Lou, Xianyuan Zhan, Qizhang Li, Wangmeng Zuo, Yang Liu, and Jingjing Liu. Adversarial contrastive learning via asymmetric infonce. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pp. 53–69. Springer, 2022.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *ArXiv*, abs/1605.07146, 2016.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *International conference on machine learning*, pp. 11278–11287. PMLR, 2020.

Shiji Zhao, Jie Yu, Zhenlong Sun, Bo Zhang, and Xingxing Wei. Enhanced accuracy and robustness via multi-teacher adversarial distillation. In *European Conference on Computer Vision*, 2022.

Dawei Zhou, Nannan Wang, Bo Han, and Tongliang Liu. Modeling adversarial noise for adversarial training. In *International Conference on Machine Learning*, pp. 27353–27366. PMLR, 2022.

Daniel Zoran, Mike Chrzanowski, Po-Sen Huang, Sven Gowal, Alex Mott, and Pushmeet Kohli. Towards robust image classification using sequential attention models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9483–9492, 2020.

A Appendix

A.1 Illustration of the asymmetric negative contrast

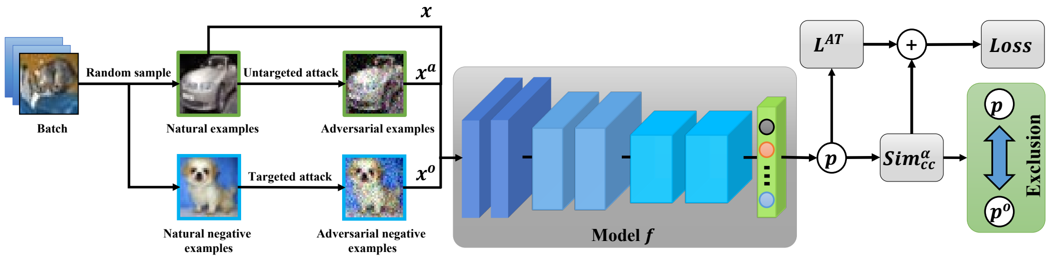


Figure 5: An illustration of the asymmetric negative contrast based on probabilities (ANC). p and p^o denote the probability vector of x and x^o, and $\mathrm{Sim}^{\alpha}(x, x^o)$ denotes the asymmetric negative contrast. Firstly, we get x^a and x^o from x by specific attacks. Then we input them to obtain their probability vectors to calculate the loss. When minimizing the loss, the similarity of p and p^o becomes smaller, causing repulsive forces to each other.

We have drawn an illustration of the asymmetric negative contrast to help readers better understand it.

A.2 EXPERIMENTS ABOUT HYPERPARAMETERS

We have used two hyperparameters in the loss function: α and ζ. α denotes the weighting factor to adjust the magnitude of the two repulsive forces, which we mentioned in Equation 3 in Section 3.2. ζ denotes the weight of the asymmetric negative contrast in the total loss, which we mentioned in Equation 1 in Section 3.2. We tune these hyperparameters on CIFAR-10 on ResNet-18.

As shown in Figure 6, there is a positive relationship between the accuracy and α. Though there is an obvious trade-off between the clean accuracy and robust accuracy when α equals from 0.5 to 1.0, we can still see an abnormal increasing trend. It is because the larger α leads to the larger repulsive force from the OE to the natural example, to prevent the natural example from being pushed into the wrong class. Besides, as shown in Figure 7, we choose $\zeta = 3.00$ in which models gain the best robust accuracy against PGD-40 in the last epoch.

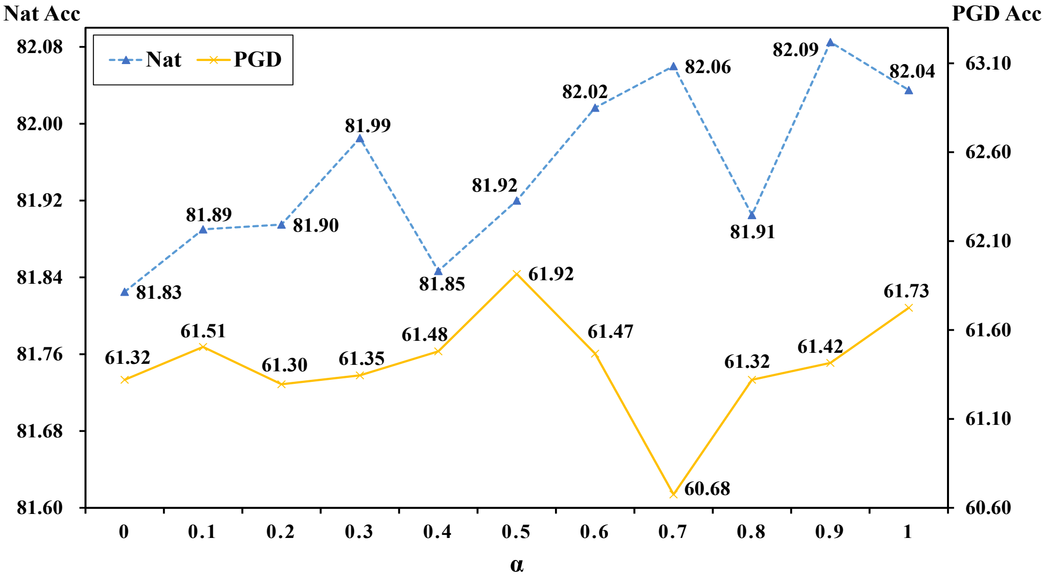


Figure 6: Clean and robust accuracy with different α.

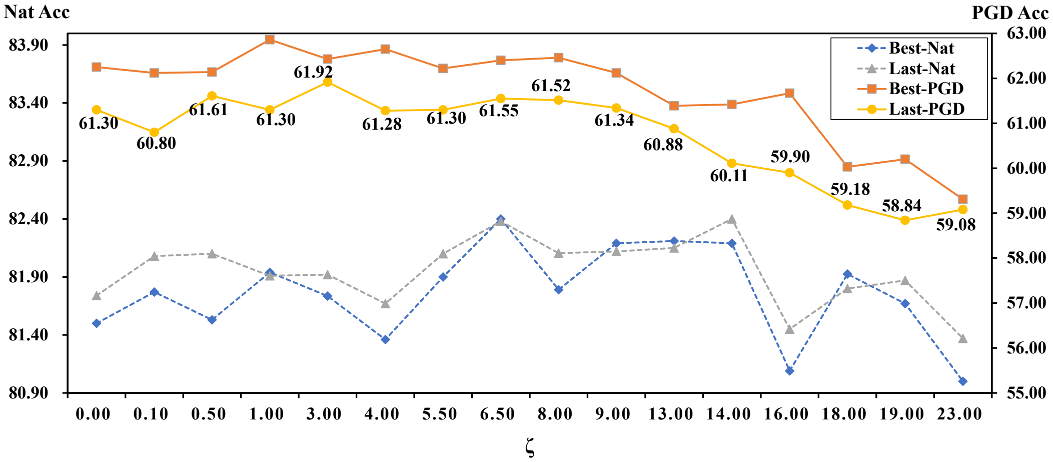


Figure 7: Clean and robust accuracy with different ζ.

A.3 FEATURE VISUALIZATION

We have drawn some frequency histograms of feature distributions on classes 1, 3 and 7 as Figure 4. As shown in Figure 8, it has similar results as the results of class 0, which indicates our methods successfully help AT to improve feature distribution and obtain robust representation.

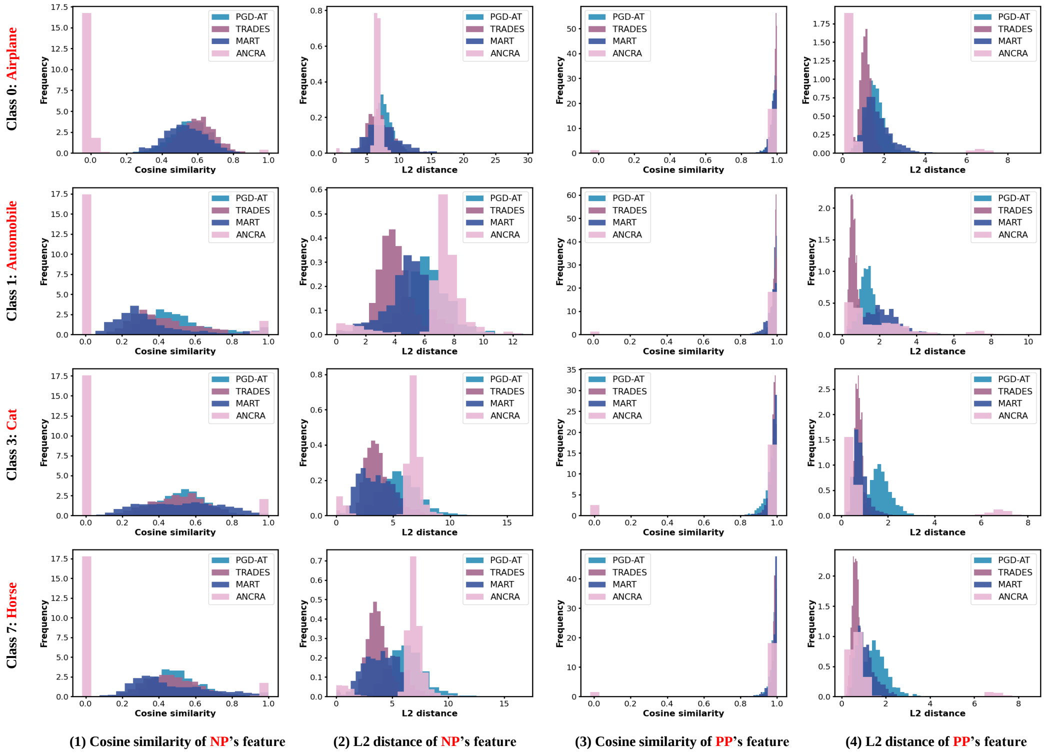


Figure 8: Frequency histograms of the L_2 distance and cosine similarity of feature of natural examples, AEs and OEs.

What's more, we have conducted several experiments of feature visualization. We use UMAP (McInnes & Healy, 2018), a visualization like t-SNE (van der Maaten & Hinton, 2008), to reduce the dimension of feature vectors and draw the distribution map. Results are shown in Figure 10 and Figure 9, where different colors denote samples of different classes. Unlike traditional AT methods, our approach can improve feature distribution by pulling close samples of the same class and pushing away samples of different classes, which follows *exclusion* and *alignment*.

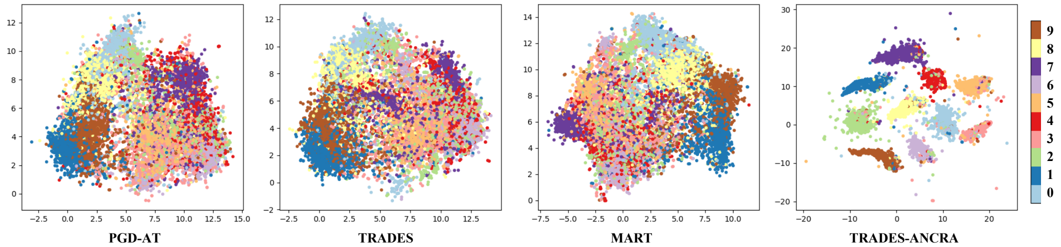


Figure 9: Feature visualization of four methods on natural and adversarial examples. Adversarial samples are crafted by PGD-10.

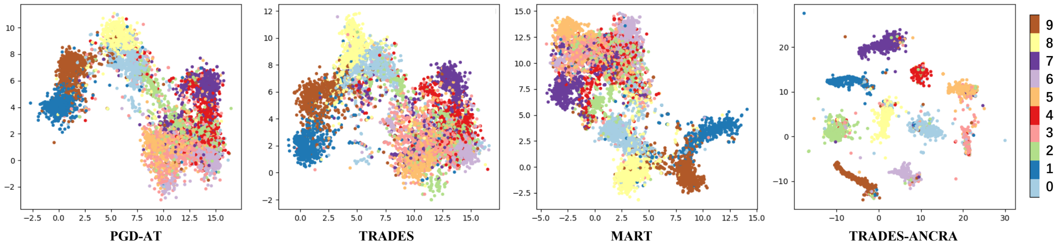


Figure 10: Feature visualization of four methods on natural examples.

A.4 EXPERIMENTS AGAINST BLACK-BOX ATTACKS

We have made some experiments against transfer-based black-box attacks. Notice that all the models are ResNet-18, so it is easy to be attacked. AEs are generated by PGD-100 on source models and tested on target models. As shown in Table 7, our method gains the best robustness among all the methods, indicating its effectiveness in the black-box scenario.

Table 7: Robustness (%) against transfer-based attacks.

Target	Source		
	PGD-AT	TRADES	MART
PGD-AT	44.73	58.25	59.65
TRADES	58.91	48.53	60.21
MART	58.66	58.46	49.26
TRADES-ANCRA	62.03	60.43	62.23

A.5 EXPERIMENTS ON LARGE DATASET

We have conducted some experiments to prove its effectiveness on large datasets. We train PreActResNet-18 (He et al., 2016b) models on Tiny-ImageNet (Deng et al., 2009). We adopt the SGD optimizer with a learning rate of 0.1, a momentum of 0.9, a weight decay of 5.0×10^{-4}, epochs of 120 and a batch size of 128. The hyperparameters are the same as our settings in the text.

Table 8: Clean and robust accuracy (%) of PreActResNet-18 trained by PGD-AT-ANCRA, TRADES-ANCRA and MART-ANCRA on Tiny-ImageNet. The error ranges are reported in brackets.

Defense	Nat	PGD-40	Adaptive PGD-40
PGD-AT	41.31(±1.2)	10.28(±0.7)	\
PGD-AT-ANCRA	43.02(±1.7)	29.79(±0.7)	11.99(±0.6)
TRADES	37.27(±0.5)	16.30(±0.8)	\
TRADES-ANCRA	38.94(±0.6)	31.24(±1.4)	17.87(±0.3)
MART	38.61(±0.9)	14.78(±0.5)	\
MART-ANCRA	43.83(±0.9)	31.44(±0.4)	13.84(±0.7)

As shown in Table 8, our methods made obvious progress in robustness compared with baselines. The performances of our method against adaptive PGD-40 are better than those against PGD-40 of baselines except MART. Besides, all the clean accuracies of ours are higher than those of baseline. These results indicate its effectiveness on big datasets.

A.6 EXPERIMENTS ON LARGE MODEL

We have conducted some experiments to prove its effectiveness on large models. We train WideResNet (Zagoruyko & Komodakis, 2016) models on CIFAR-10. We adopt the SGD optimizer with a learning rate of 0.1, a momentum of 0.9, a weight decay of 2.0×10^{-4}, epochs of 76 and a batch size of 128. α, β are the same as our settings in the text. $\zeta = 6.0$ when training WideResNet-28-10 and $\zeta = 3.0$ when training WideResNet-34-10.

Table 9: Clean and robust accuracy (%) of WideResNet models trained by TRADES-ANCRA.

Defense	Model	Nat	PGD	Adaptive PGD	AA	Adaptive AA
TRADES	WideResNet-28-10	82.47	57.08	\	51.11	\
TRADES-ANCRA	WideResNet-28-10	83.61	78.82	58.60	65.87	51.85
TRADES	WideResNet-34-10	82.04	56.47	\	50.79	\
TRADES-ANCRA	WideResNet-34-10	83.61	79.31	59.35	66.28	51.99

As shown in Table 9, our method has made some enhancements in clean and robust accuracies. Our method has better natural accuracies than baselines by 1.14% and 1.57%. Besides, the accuracies of our method against adaptive attacks are higher than those of baselines against vanilla attacks (e.g., 51.99%>50.79%). These results indicate its effectiveness on large models.

A.7 Comparative experiments with methods in the RobustBench

Table 10: Comparative experiments with methods in the RobustBench. All the models are in ResNet-18 trained on CIFAR-10. AA denotes robust accuracy against AutoAttack. Best results are in **bold**.

Defense	Nat	AA
Sehwag et al. (2021)	**87.35**	58.50
Addepalli et al. (2022b)	85.71	52.48
Addepalli et al. (2022a)	80.24	51.06
PGD-AT-ANCRA	85.10	59.15
TRADES-ANCRA	81.70	**59.70**
MART-ANCRA	84.88	59.62

We have made a comparison with the current state-of-the-art performances listed on the RobustBench[3] on ResNet-18. The results are shown in Table 10. Compared with those methods without synthetic or extra data (i.e., Addepalli et al. (2022b) and Addepalli et al. (2022a)), our method has a higher robust accuracy than theirs by 7.0%. And our method has even outperformed the methods with synthetic data (Sehwag et al., 2021) in robustness. Though the clean accuracy of Sehwag et al. (2021) is more than ours by 5.6%-2.2%, the best robust performance has indicated the effectiveness of our methods. Experiment results in ResNet-18 have shown our superiority of robustness.

A.8 Time cost

In our experiments, our TRADES-ANCRA only costs more time than TRADES by 3.1 hours (3.1=9.3-6.2) in 120 epochs. Considering the significant gain in clean and robust accuracy resulting from the proposed method, the cost is relatively worthwhile.

[3] https://robustbench.github.io/