Phase-aware Adversarial Defense for Improving Adversarial Robustness

Dawei Zhou [1] **Nannan Wang** [1] **Heng Yang** [2] **Xinbo Gao** [3] **Tongliang Liu** [4 5]

Abstract

Deep neural networks have been found to be vulnerable to adversarial noise. Recent works show that exploring the impact of adversarial noise on intrinsic components of data can help improve adversarial robustness. However, the pattern closely related to human perception has not been deeply studied. In this paper, inspired by the cognitive science, we investigate the interference of adversarial noise from the perspective of image phase, and find ordinarily-trained models lack enough robustness against phase-level perturbations. Motivated by this, we propose a joint adversarial defense method: a *phase-level adversarial training mechanism* to enhance the adversarial robustness on the phase pattern; an *amplitude-based pre-processing operation* to mitigate the adversarial perturbation in the amplitude pattern. Experimental results show that the proposed method can significantly improve the robust accuracy against multiple attacks and even adaptive attacks. In addition, ablation studies demonstrate the effectiveness of our defense strategy.

1. Introduction

Many studies have demonstrated that deep neural networks (DNNs) are easily fooled by imperceptible but misleading perturbations, *i.e,* carefully crafted adversarial noise (Goodfellow et al., 2015; Szegedy et al., 2014; Ma et al., 2018; Zhang et al., 2019; Croce & Hein, 2020b; Wu et al., 2020; Yu et al., 2022b). This vulnerable behavior has seriously threatens many decision-critical deep learning applications (LeCun et al., 1998; He et al., 2016; Zagoruyko & Komodakis, 2016; Simonyan & Zisserman, 2015; Kaiming et al., 2017; Ma et al., 2021), and thus may lead to a lack of trustworthiness in deep learning.

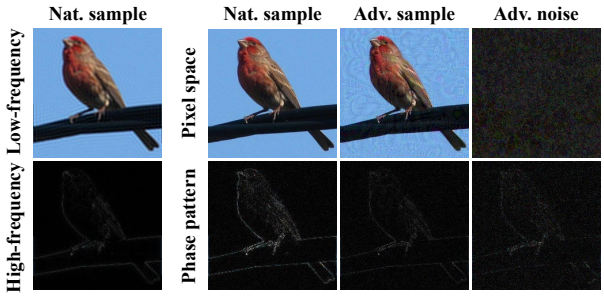

Figure 1. The illustrations of the low-frequency component, high-frequency component and phase pattern. Nat. and Adv. samples denote natural and adversarial samples. The noise is crafted by PGD attack (Madry et al., 2018).

To alleviate the vulnerability, in addition to improving regularization terms in loss functions (Zhang et al., 2019; Wang et al., 2019), recent works study the impact of adversarial noise on the different intrinsic components of the data. The work in Tsipras et al. (2018) explores the different distributions of adversarial noise on the background and objects. Some works (Yin et al., 2019; Wang et al., 2020; Olivier et al., 2021; Song & Deng, 2021) analyze the destructiveness of perturbations in the high-frequency component or the low-frequency component, respectively. They then guide the model to focus on more robust components to enhance the adversarial robustness at the source. However, these components usually do not sufficiently and explicitly reflect structural information (see Figure. 1), and are not perceivable to humans (Wang et al., 2020; Chen et al., 2021). The studies in (Biederman, 1987; Landau et al., 1998; Geirhos et al., 2018; Samuelson & Smith, 2005; Schmidt et al., 2020) show that the human vision mainly relies on structural semantics to robustly distinguish objects. We thus expect to mine a component that is closely related to human perception to help defend against adversarial noise.

The image signal in the pixel space can be converted into the frequency signal by Fourier transform, and further decou-

[1]School of Telecommunications Engineering, State Key Laboratory of Integrated Services Networks, Xidian University, Xian, Shaanxi, China [2]Shenzhen AiMall Tech, Shenzhen, China [3]Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing, China [4]Mohamed bin Zayed University of Artificial Intelligence, Masdar City, Abu Dhabi, United Arab Emirates [5]University of Sydney, Darlington, NSW, Australia. Correspondence to: Nannan Wang <nnwang@xidian.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

pled into the amplitude spectrum and the phase spectrum. The amplitude spectrum carries the pixel intensity information. The phase spectrum can reflect the highly informative structural features (Ghiglia & Pritt, 1998; Morrone et al., 1986; Morrone & Owens, 1987; Kovesi, 2000; Zhang et al., 2011) (see Figure. 1), which is consistent with the need of the human perception. In addition, the psychophysical and neuroscience evidences (Pollen & Ronner, 1981; 1983; Concetta Morrone & Burr, 1988; Freeman & Simoncelli, 2011; Zhang et al., 2014; Gladilin & Eils, 2015) suggest that humans tend to leverage more phase information to understand and recognize objects. Based on this, we rethink the impact of adversarial noise from the phase perspective.

We first visually observe the adversarial perturbation in the phase pattern by constructing samples containing only phase pattern. As shown in Fig. 1, adversarial noise can perturb or eliminate some structural semantics of the objective, *e.g.*, the bird's head. In this case, ordinarily-trained models may fail to extract sufficient features for making correct predictions. Therefore, we believe that the *adversarial noise can cause severe perturbations to the phase pattern* and speculate that these perturbations are important negative factors leading to the degradation of performance.

To validate the above surmise, proof-of-concept studies are conducted (see Section. 3). They show that phase-level perturbations severely degrade the classification accuracy, even more destructively than amplitude-level perturbations. If the phase-level robustness can be enhanced, the model is expected to be less vulnerable to adversarial noise. In addition, these studies present that not all phase features are conducive to adversarial robustness. We argue that the models based on the ordinary training manner may *pay biased attention to predictive but easily perturbed phase features, but ignore pivotal semantics*. These phenomena indicate that defending adversarial noise on the phase pattern is beneficial in help models achieve more robust performances.

Motivated by above studies, we propose an adversarial defense method from the perspective of phase (see Section. 4.2.1). Specifically, considering that the ordinary training manner lack an explicit guidance on the phase pattern, we design a *phase-level adversarial training* mechanism. To enforce the model to mine and learn robust features over the phase pattern, this mechanism leverages the amplitude spectrum of natural data to replace that of the adversarial data. It recombines the clean amplitude spectrum with the adversarial phase spectrum to obtain new training data. This strategy can help the model further leverage key phase features to make predictions and learn a more robust decision boundary against the adversarial noise (see Section. 5.2), which may be consistent with human perception (Chen et al., 2021; Zhang et al., 2011).

In the inference stage, the input adversarial sample does not have a clean amplitude spectrum. This results in the threat of perturbations remain in the amplitude pattern (even if the model is trained to focus on the phase pattern). To alleviate this problem, a straightforward approach is to take a natural sample as a reference sample and swap its amplitude spectrum with that of the input sample. Unfortunately, this approach is easy to cause obfuscated information in the recombined images, which affects the normal prediction of the model, and even interferes with human perception (see Section. 4.2.2). We therefore design an *amplitude-based pre-processing operation* that leverages the style transfer technique to construct an transitional reference sample as the new reference. In this way, the recombined sample has less obfuscation and amplitude-level perturbations.

Considering the overall effect of the defense, we combine the adversarial training and pre-processing operation to jointly optimize model parameters for optimal performance. Experimental results show that our method can provide significant gains in the robust accuracy.

The main contributions are summarized as follows:

- We qualitatively and quantitatively investigate the impact of adversarial noise from the perspective of human perception-related phase. We find that ordinarily-trained models lack enough adversarial robustness against phase-level perturbations.
- We propose a *Phase-aware Adversarial Defense* (PAD) method to alleviate the vulnerability of models. The method aims to enhance the adversarial robustness on the phase pattern and help attenuate the interference of amplitude-level perturbations.
- Experimental results show that our method could effectively improve adversarial robustness against multiple adversarial attacks, including adaptive attacks. Ablation studies are performed to demonstrate the effectiveness of each module.

2. Related Work

Adversarial attacks. Existing studies have proposed a variety of adversarial attacks. For example, gradient-based attacks have projected gradient descent (PGD) attack (the strongest first-order attack) (Madry et al., 2018), autoattack (AA) (Croce & Hein, 2020b), variance tuning momentum iterative fast gradient sign method (VMI-FGSM) (Wang & He, 2021) and optimization-based attacks: fast adaptive boundary (FAB) attack (Croce & Hein, 2020a), Carlini&Wagner (C&W) attack (Carlini & Wagner, 2017b). Moreover, some attacks, such as the spatial transform attack (STA) (Xiao et al., 2018) aims to perturb the spatial structure information of the objective. In addition to above attacks, we examine the defense by using an expectation

over transformation (EOT) attack (Athalye et al., 2018b) and a backward pass differentiable approximation (BPDA) attack (Athalye et al., 2018a).

Adversarial defenses. The adversarial noise promotes the development of adversarial defenses. A representative defense strategy is devoted to enhancing the adversarial robustness of models in an adversarial training manner (Madry et al., 2018; Ding et al., 2019; Zhang et al., 2019; Wang et al., 2019; Wu et al., 2020; Yu et al., 2022a; Zhou et al., 2022; Wang et al., 2022; Zhang et al., 2022; Clarysse et al., 2022; Carlini et al., 2022; Li et al., 2023). In addition, the pre-processing based defense strategy has also been extensively studied. This strategy typically aims to remove adversarial noise by learning denoising maps (Liao et al., 2018; Naseer et al., 2020; Zhou et al., 2021a) or feature-squeezing functions (Guo et al., 2018b). The work in Ilyas et al. (2019) constructed robust training set to guide the model to learn robust features and achieve robust performances. The works in Yin et al. (2019); Wang et al. (2020); Olivier et al. (2021); Zhou et al. (2021b); Song & Deng (2021) analyzed adversarial noise in the high-frequency or low-frequency component and devised targeted methods to enhance robustness. Differently, our work designs adversarial training from a phase perspective. The phase pattern is closely consistent with structural information. More details can be found in Appendix. A.

3. Adversarial Perturbation on Phase Pattern

Studying the impact of adversarial noise from a phase perspective is considered to be beneficial for enhancing adversarial robustness. This is because human vision mainly relies on semantic information to robustly identify the objective (Biederman, 1987; Landau et al., 1998; Geirhos et al., 2018; Samuelson & Smith, 2005; Schmidt et al., 2020). The phase pattern of an image is closely related to the structural information of the image. The phase is a description of the position of the signal. The phase spectrum carries the position information of different parts, that is, it can reflect the outline and structural information of the object in the image (Pollen & Ronner, 1981; 1983; Ghiglia & Pritt, 1998; Kovesi, 2000; Zhang et al., 2011; Freeman & Simoncelli, 2011). Some studies (Oppenheim & Lim, 1981; Concetta Morrone & Burr, 1988; Zhang et al., 2014; Gladilin & Eils, 2015) also show that humans tend to leverage more phase information to understand and recognize object. If we can understand whether the noise has serious perturbations to the phase pattern, and clear the damage degree of these perturbations to model performances, we can infer a potential cause for the vulnerability and design targeted adversarial defense. Therefore, we conduct intuitive observations as well as proof-of-concept studies.

Qualitative study. We first concretize the phase pattern

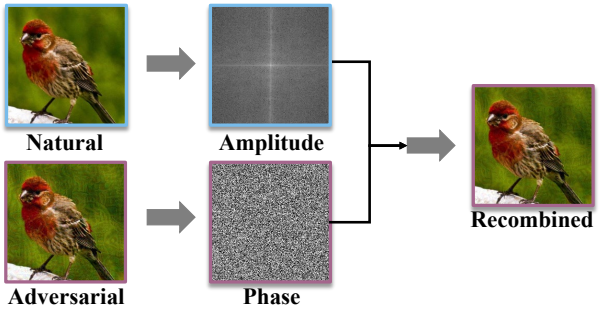

Figure 2. Schematic representation of replacing the natural phase spectrum with the adversarial phase spectrum. We obtain the amplitude spectrum of a natural sample and the phase spectrum of an adversarial sample, and perform inverse Fourier Transformation on them. The similar operation is conducted for replacing the natural amplitude spectrum with the adversarial amplitude spectrum.

of the image by performing the inverse Fourier transform on only the phase spectrum (replacing the amplitude spectrum with a constant matrix). As shown in Figure. 1 and Appendix. B, adversarial noise perturb or eliminate some structural semantics of the objective. For example, the contour of the bird's head is significantly faded. We note that these features are often not the core information for identifying birds or snakes, as we humans can still clearly recognize them via the beak, claws or overall contour features. They may be predictable but easily disturbed phase features. The observation leads us to believe that *adversarial noise causes malicious perturbations to the phase patterns*, and speculate that these perturbations play an important role in the vulnerability of the deep learning model.

Quantitative study. To validate the above speculation, we perform a quantitative study. Specifically, as shown in Figure. 2, we utilize the phase spectrum and amplitude spectrum of the adversarial sample to replace the corresponding spectrum of the natural sample respectively, and then recombine the adversarial spectrum and the natural spectrum via the inverse Fourier transform. We next input the recombined samples into an ordinarily-trained model and calculate the classification accuracy.

As shown in Figure. 3, we find that the phase-level perturbation significantly reduces the classification accuracy, and its reduction is greater than that caused by the amplitude-level perturbation. As the number of attack steps increases, phase-level perturbations continue to cause damage, even if amplitude-level perturbations no longer degrade the accuracy. This shows that the *phase-level perturbation is indeed an important factor causing model vulnerability*, and it is more destructive to the model than the amplitude-level perturbation. As the attack deepens, the attack is able to further find weak points in the phase pattern. Considering

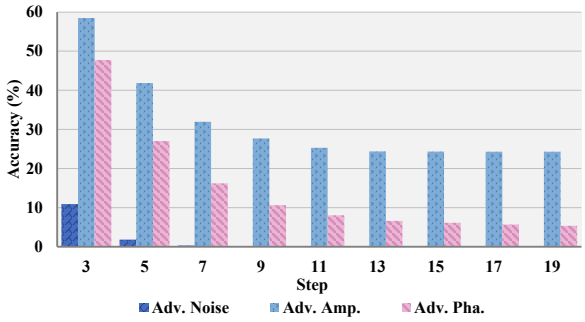

Figure 3. The impact of phase-level and amplitude-level perturbations on the classification accuracy. *Adv. Noise* indicates that the sample contains normal adversarial noise. *Adv. Amp./Pha.* indicates that only the phase/amplitude spectrum contains adversarial perturbations. *Step* denotes the number of attack steps. The adversarial noise are crafted by using PGD (Madry et al., 2018) with perturbation budget 8/255 and step size 1/255. The dataset and network architecture are *CIFAR-10* and ResNet-18.

Table 1. The pixel-level noise norm caused by phase-level perturbations and amplitude-level perturbations on *CIFAR-10*. 'Original' denotes the original adversarial sample, 'Amplitude' denotes the adversarial sample with only amplitude-level perturbations and 'Phase' denotes the adversarial sample with only phase-level perturbations. The step size of PGD attack is 15.

	Original	Amplitude	Phase
Accuracy	0.01%	24.32%	6.12%
Accuracy$_{com}$	0.01%	24.34%	6.12%
ℓ_∞	0.0314	0.0515	0.0374
ℓ_2	1.3860	0.9931	0.9385

the above visual and statistical results together, we argue that *the model lacks enough robustness on the phase pattern due to its insufficient attention to core phase features*, which is rarely mentioned in previous works.

In addition, considering that lower accuracy caused by phase-level perturbations may due to the larger noise size, we present the noise norm at an attack step of 15 in Table. 1. 'Original' denotes the original adversarial sample, 'Amplitude' denotes the adversarial sample with only amplitude-level perturbations and 'Phase' denotes the adversarial sample with only phase-level perturbations. It can be seen that phase-level perturbations lead to less noise norms while having stronger attack effects (*i.e.*, lower 'Accuracy'), which indicates that phase-level perturbations do have greater impacts on the performance of the target model than the amplitude-level perturbations. We note that the ℓ_∞ norms of 'Amplitude' and 'Phase' are larger than that of 'Original'. This is because ℓ_∞ norm calculates the maximum difference

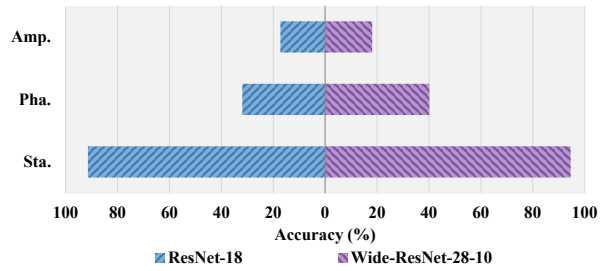

Figure 4. The consistency between model's predictions and the phase/amplitude pattern on *CIFAR-10*. *Sta.* denotes the standard accuracy. *Pha./Amp.* mean to use the label of the sample corresponding to the phase/amplitude pattern as the ground-truth label. The results show that the model is more dependent on the phase than the amplitude, but the consistency is still insufficient.

of pixel intensity, and there may be individual pixels that vary significantly during the recombination. Fortunately, as can be seen from the values of ℓ_2 norm, the perturbation degrees of 'Amplitude' and 'Phase' are significantly lower than that of 'Original'. In addition, we calculate the accuracy ('Accuracy$_{com}$') for the recombined adversarial samples restricted within a ℓ_∞ of $8/255$. The new results are almost identical to the original results. These results show that the phase-level perturbations play important roles in misleading the target model.

Furthermore, we investigate the consistency between the model's predictions and the phase information. Given a set of test samples, we randomly swap their amplitude spectrum, and then feed the recombined samples into the model. The result in Figure. 4 shows that the prediction results are not consistent with the structural information in the phase pattern (only about 30% and 40% accuracy). The models do not rely closely on the phase pattern to make predictions. This counter-intuitive phenomenon does not fully fit with human perception. The investigation again suggests that the ordinarily-trained model do not (or cannot) stably capture key structural semantics in the phase pattern for classification, which will naturally lead to its sensitivity to amplitude variations and phase-level perturbations.

Based on our qualitative and quantitative studies, we hope to find a defense that can enforce the model to focus on core features in the phase pattern. Defending against adversarial noise from a phase perspective is expected to effectively help models achieve more robust performances, as confirmed by empirical results in Section. 5.1.

4. Methodology

In this section, we first describe some preliminaries about notation and the problem setting. Then, we introduce the

composition of the proposed method and present its algorithm procedure.

4.1. Preliminary

Notation. We use *capital* letters such as X and Y to denote random variables, and *lower-case* letters such as x and y to denote instances of random variables X and Y, respectively. For norms, we use $\|x\|$ to denote a generic norm. Specific examples of norms include $\|x\|_\infty$, the L_∞-norm of x, and $\|x\|_2$, the L_2-norm of x. Let $\mathbb{B}(x, \epsilon)$ represent the neighborhood of x: $\{\tilde{x} : \|\tilde{x} - x\| \leq \epsilon\}$, where ϵ is the perturbation budget. We define the *classification function* as $f : \mathcal{X} \to \{1, 2, \ldots, C\}$, where $\mathcal{X}$ is the feature space of X. The function can be parameterized by neural networks.

Problem setting. This paper mainly focuses on the issue of adversarial robustness in classification tasks. Let X and Y be the variables for natural samples and natural labels (*i.e.,* the ground truth labels of natural samples) respectively. We sample natural examples $\{(x_i, y_i)\}_{i=1}^n$ according to the distribution of the variables (X, Y), where $(X, Y) \in \mathcal{X} \times \{1, 2, \ldots, C\}$. Given a natural example (x, y) and a classifier f parameterized by a deep learning model h_θ with the model parameter θ, the adversarial sample x' satisfies one of the following constraints:

$$f(x') \neq y \quad \text{s.t.} \|x - x'\| \leq \epsilon. \tag{1}$$

In this paper, our aim is to design an adversarial defense to improve the robustness of the model from the phase perspective, which is closely related to the human perception. Since our work is mainly concerned with adversarial attacks and defenses on images, we thus utilize the discrete Fourier transforms (DFT). We denote DFT and the inverse discrete Fourier transform (IDFT) by $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot, \cdot)$. In addition, we utilize ϕ_x, ξ_x to denote the phase spectrum and amplitude spectrum of x, and utilize $\phi_{x'}$, $\xi_{x'}$ to denote the phase spectrum and amplitude spectrum of x'. The procedure to obtain the phase and amplitude spectra is denoted as $\phi_x = \mathcal{F}_\phi(x)$ and $\xi_x = \mathcal{F}_\xi(x)$. Similarly, the procedure to recover sample from its phase spectrum and amplitude spectrum is denoted by $x = \mathcal{F}^{-1}(\phi_x, \xi_x)$. The more details on the discrete Fourier transform can be found in Appendix. C.

4.2. Phase-aware adversarial defense

In this paper, we are committed to defending against adversarial noise from a image phase perspective. We propose an combined *Phase-aware Adversarial Defense* (PAD) method. This method consists of a phase-level adversarial training and an amplitude-based pre-processing mechanism. A joint optimization strategy is provided to achieve the optimal overall performance.

4.2.1. PHASE-LEVEL ADVERSARIAL TRAINING

Based on our explorations, we note that the phase pattern contains explicit structural information. Cognitive science researches have demonstrated that the phase pattern play a crucial role in the process of human understanding and recognizing objectives (Pollen & Ronner, 1981; 1983; Concetta Morrone & Burr, 1988; Freeman & Simoncelli, 2011; Zhang et al., 2014; Gladilin & Eils, 2015). However, the ordinary training and standard adversarial training strategies lack explicit phase-level guidance. They thus may fail to sufficiently and effectively mine phase features.

To address this issue, we plan to enforce the model to devote itself to extracting and learning pivotal phase features for enhancing its adversarial robustness against phase-level perturbations. Given a natural sample x and its adversarial sample x', we obtain the adversarial phase spectrum $\phi_{x'}$ of x' and the natural amplitude spectrum ξ_x of x via DFT, respectively. Then, we recombine them and generate recombined sample $\mathcal{F}^{-1}(\phi_{x'}, \xi_x)$ via IDFT. We exploit the recombined sample to construct a classification loss:

$$\mathcal{L}_c(x, x', y) = -\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{y_i} \cdot \log(h_\theta(\mathcal{F}^{-1}(\phi_{x'_i}, \xi_{x_i}))), \tag{2}$$

where N denotes the number of samples, h_θ denotes the target model and $\boldsymbol{y}$ denotes the one-hot label. This loss can promote the model to autonomously learn robust decision boundaries on the phase pattern.

In addition, we note that there are still residual structural semantics in the phase pattern of adversarial examples, which may be difficult to be disturbed by adversarial noise. Moreover, in human vision, we can recognize the object via these residual semantics, which means that they may be pivotal phase features. Of course, this phase pattern also contains a lot of adversarial perturbations, and using it directly may not be a good choice. Therefore, we can jointly utilize the natural phase pattern and the adversarial phase pattern to guide the model to learn their shared features, *i.e.*, the pivotal phase features. A similarity loss is formulated as:

$$\mathcal{L}_s(x, x') = \frac{1}{N} \sum_{i=1}^{N} \ell_d(h_\theta(\mathcal{F}^{-1}(\phi_{x'_i}, \xi_{x_i})), h_\theta(x_i)), \tag{3}$$

where ℓ_d denotes the distance metric, we use the Kullback-Leibler divergence here. The optimization objective for the target model h_θ is as follows:

$$\arg\min_{\theta \in \Theta} \mathcal{L}_c(x, x', y; \theta) + \alpha \cdot \mathcal{L}_s(x, x'; \theta), \tag{4}$$

where Θ is the set of model parameters and α is a hyper-parameter to tune the weights of these two terms.

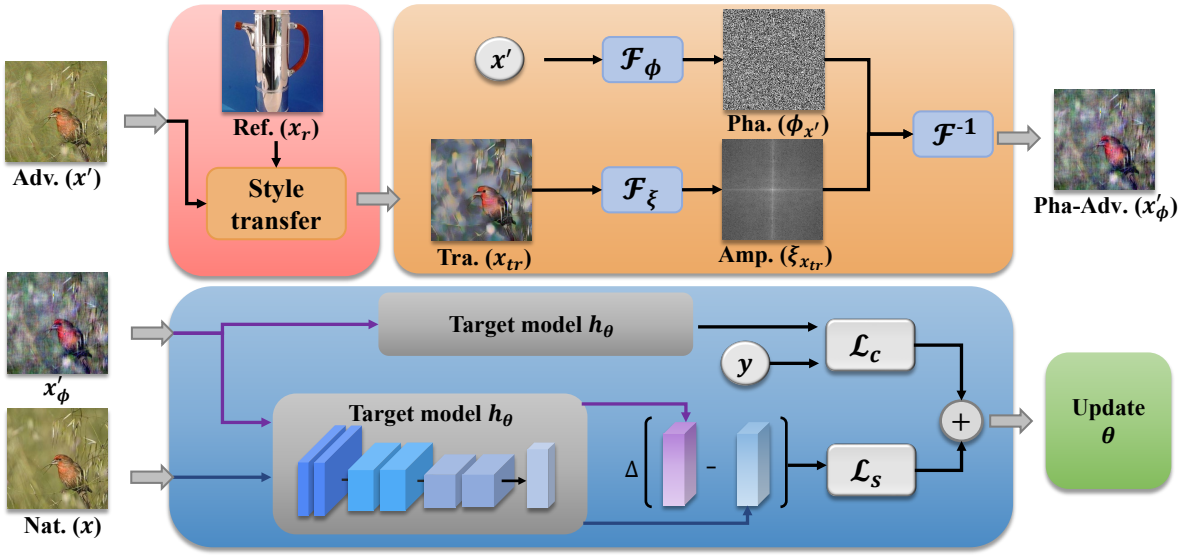

Figure 5. The training procedure of the phase-aware adversarial defense. *Adv.*, *Nat.*, *Ref.*, *Tra.*, *Pha.*, *Amp.* and *Pha-Adv.* mean the adversarial sample, natural sample, reference sample, transitional reference sample, phase spectrum, amplitude spectrum and recombined phase-level adversarial sample, respectively. Δ denotes the distance metric used in Equation. 3. The pre-processing procedure of our method in the inference stage is similar to the pink and orange parts.

4.2.2. AMPLITUDE-BASED PRE-PROCESSING

The phase-level adversarial training strategy replaces the amplitude spectrum of the adversarial sample with that of the natural sample during the training process. However, the threat of the adversarial perturbation on the amplitude pattern remains in the inference stage. To decrease this risk for further prompting the robust accuracy, we design a amplitude-based pre-processing mechanism.

A simple approach is to utilize an additional natural sample as a reference sample, and use its amplitude spectrum to replace that of the test sample. This per-processing procedure can be formulated as:

$$\hat{x}_t = \mathcal{F}^{-1}(\mathcal{F}_\phi(x_t), \mathcal{F}_\xi(x_r)), \tag{5}$$

where x_t denotes the test sample and x_r denotes the natural reference sample. The generated sample $\hat{x}_e$ is then input into the target model for prediction. This procedure needs to be performed for all input samples as it is unknown whether the input samples are malicious.

Unfortunately, this approach may cause some recombined samples to suffer from obfuscated information, which affects the normal prediction of the model. These obfuscations even interfere with human perception (See Figure. 6). This indicates that using a random sample directly as a reference is not an optimal choice. The matching of the amplitude spectrum used for replacement with the original phase spectrum is an important factor to be considered. We therefore plan to build new reference samples.

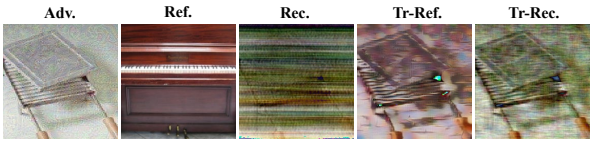

Figure 6. The illustration of recombined samples. *Adv.*, *Ref.*, *Rec.*, *Tr-Ref.* and *Tr-Rec.* denote the adversarial sample, original reference sample, the recombined sample with *Ref.*, the transitional reference sample and the recombined sample with *Tr-Ref.*. The obfuscated information in *Tr-Rec.* is significantly less than that in *Rec.* and the objective can be clearly seen in *Tr-Rec.*. More examples can be found in Appendix. E.

The amplitude spectrum mainly represents the pixel intensity distribution, which is closely related to the style texture (Oppenheim & Lim, 1981; Tolhurst et al., 1992; Hansen & Loschky, 2013; Zibar et al., 2015). Based on this, we leverage the style transfer technology to overwrite the style of the input sample with the natural style of the reference sample, and generate an transitional reference sample as the new reference. This transitional reference sample has a similar (but not identical) phase pattern to the test sample, while having natural amplitude information. It can thus match well with the phase spectrum of the test sample and facilitate the recombined sample to reduce obfuscated information. Empirical results in Section. 5.3 demonstrate its

effectiveness. The pre-processing procedure in Equation. 5 is thus reformulated as:

$$\hat{x}_t = \mathcal{F}^{-1}(\mathcal{F}_\phi(x_t), \mathcal{F}_\xi(\mathcal{T}(x_t, x_r))), \tag{6}$$

where $\mathcal{T}(x_t, x_r)$ means to transfer the style of x_r to x_t. We utilize the classic *AdaIN* (Huang & Belongie, 2017) to perform the style transfer.

4.2.3. JOINT OPTIMIZATION

To improve the overall effectiveness of our combined defense, we incorporate the transitional reference sample used in the pre-processing mechanism into the training process. This can enable the target model to adapt to the data distribution of the recombined samples, resulting in better natural and robust performance. The schematic diagram of the proposed method is shown in Figure. 5. The loss functions in Equation. 2 and 3 are reformulated as:

$$\mathcal{L}_c(x_r, x', y; \theta) = \frac{1}{N}\sum_{i=1}^{N}[\boldsymbol{y_i} \cdot \log(h_\theta(\zeta(x'_i, x_{r_i}))), \tag{7}$$

$$\mathcal{L}_s(x, x', x_r; \theta) = \frac{1}{N}\sum_{i=1}^{N}\ell_d(h_\theta(\zeta(x'_i, x_{r_i})), h_\theta(x_i)), \tag{8}$$

where $\zeta(x'_i, x_r)$=$\mathcal{F}^{-1}(\phi_{x'}, \mathcal{F}_\xi(\mathcal{T}(x', x_r)))$ and x_r denotes the reference sample. The definitions of other symbols are the same as in Equation. 7. The optimization objective is reformulated as:

$$\min_{\theta\in\Theta} \mathcal{L}_c(x_r, x', y; \theta) + \alpha \cdot \mathcal{L}_s(x, x', x_r; \theta). \tag{9}$$

The overall training procedure is presented in Algorithm. 1. Specifically, for each mini-batch natural samples $\mathcal{B}$=$\{x_i\}_{i=1}^n$ sampled from natural training set, we first generate adversarial samples $\{x'_i\}_{i=1}^n$ via a strong attack algorithm. At the same time, we randomly select samples different from $\mathcal{B}$ from the natural training set as the reference set $\mathcal{R}$=$\{x_{r_i}\}_{i=1}^n$. We then construct the transitional reference sample x_{tr}=$\mathcal{T}(x', x_r)$ and the phase-level adversarial sample x'_ϕ=$\mathcal{F}^{-1}(\phi_{x'}, \mathcal{F}_\xi(x_{tr}))$. Next, we compute the classification loss via Equation. 7 and the similarity loss via Equation. 8. Finally, we update the model parameter θ via Equation. 9. By iteratively conducting the procedures of generating adversarial samples and training the models, the model parameters are expected to be adversarially optimized. The code can be found in https://github.com/dwDavidxd/PAD.

Algorithm 1 Phase-aware Adversarial Defense (PAD).

Input: Target model h_θ, batch size n, perturbation budget ϵ and training set $\mathcal{D}$.

repeat
  Obtain mini-batch natural samples $\mathcal{B} = \{x_i\}_{i=1}^n$ and reference samples $\mathcal{R} = \{x_{r_i}\}_{i=1}^n$ from $\mathcal{D}$;
  for $i = 1$ to n (in parallel) **do**
    Craft adversarial sample x'_i at the given perturbation budget ϵ against x_i;
    Construct the transitional reference sample $x_{tr_i} = \mathcal{T}(x'_i, x_{r_i})$;
    Obtain the phase-level adversarial sample $x'_\phi = \mathcal{F}^{-1}(\phi_{x'_i}, \mathcal{F}_\xi(x_{tr_i}))$;
    Compute the classification loss $\mathcal{L}_c$ via Eq. 7;
    Compute the similarity loss $\mathcal{L}_s$ via Eq. 8;
  end for
  Back-pass the gradients and update θ via Eq. 9;
until training converged.

Output: Model parameter θ.

5. Experiment

5.1. Experiment setup

Dataset. We use two classic datasets *CIFAR-10* (Krizhevsky et al., 2009) and *Mini-ImageNet* (Vinyals et al., 2016) to evaluate the effectiveness of our method. All images are normalized into [0,1], and are performed data augmentations in the training stage. we utilize ResNet-18 (He et al., 2016) and Wide-ResNet (WRN-28-10) (Zagoruyko & Komodakis, 2016) as target models for *CIFAR-10* and *Mini-ImageNet*, respectively.

Attack settings. We use seven types of adversarial attack algorithms to evaluate the performances of defenses. They are L_∞-norm PGD (Madry et al., 2018), L_∞-norm AA (Croce & Hein, 2020b), L_∞-norm VMIFGSM(Wang & He, 2021), L_∞-norm EOT-PGD (Athalye et al., 2018b), L_∞-norm FAB(Croce & Hein, 2020a), L_2-norm CW (Carlini & Wagner, 2017b) and STA (Xiao et al., 2018). The iteration numbers for PGD, EOT-PGD, VMIFGSM and FAB are set to 20, and those for STA and CW are 10 and 200, respectively. For *CIFAR-10* and *Mini-ImageNet*, the perturbation budget ϵ for L_∞-norm attacks is set to $8/255$.

Defense settings. We use three classic adversarial training methods as the baselines: standard adversarial training (AT) (Madry et al., 2018), optimized adversarial training methods TRADES (Zhang et al., 2019) and MART (Wang et al., 2019). For all defenses, we use the L_∞-norm non-target PGD-10 to craft adversarial noise in the training stage. All the defenses are trained using SGD (Andrew & Gao, 2007) with momentum 0.9. For adversarial training, the initial learning rate is set to 2×10^{-1} corresponding to the batch size 256 according to work in Pang et al. (2020), and is divided by 10 at the 75-th and 90-th epoch. The weight decay is 2×10^{-4} for *CIFAR-10*, and is 5×10^{-4} for *Mini-ImageNet*. The epoch number is set to 91 by using the early-stopping strategy (Rice et al., 2020). More details can be found in Appendix. D.

Table 2. Robust accuracy (percentage) of defense methods against adversarial attacks on *CIFAR-10* and *Mini-ImageNet*. We show the most successful defense with **bold**.

CIFAR-10 (ResNet-18)				
Defense	AT	TRADES	MART	PAD
None	83.34	80.71	79.38	**83.56**
PGD	47.23	50.97	52.83	**58.24**
AA	44.45	47.36	46.32	**57.07**
VMIFGSM	47.15	50.90	52.71	**58.12**
EOT-PGD	47.06	50.84	52.69	**57.98**
FAB	46.28	48.56	47.47	**74.90**
CW	12.25	40.33	33.88	**75.06**
STA	0.59	3.19	2.72	**56.61**
Mini-ImageNet (WRN-28-10)				
Defense	AT	TRADES	MART	PAD
None	**50.76**	49.92	47.88	50.63
PGD	24.50	25.55	25.86	**30.04**
AA	18.11	19.17	19.02	**24.79**
VMIFGSM	24.57	25.21	25.46	**29.91**
EOT-PGD	24.53	25.15	25.34	**29.83**
FAB	20.05	21.76	20.69	**37.62**
CW	32.55	35.10	34.27	**46.67**
STA	0.21	1.67	1.34	**23.15**

Table 3. Robust accuracy (percentage) of defense methods against adaptive adversarial attacks on *CIFAR-10*.

Attack	None	PGD-10	PGD-20	BPDA
APE-G	81.63	35.19	32.37	31.60
HGD	82.05	40.93	38.26	37.52
PAD	**82.83**	**55.32**	**54.91**	**54.33**

5.2. Robustness evaluation

Defending against general attacks. We first evaluate the effectiveness of the proposed method against general adversarial attacks. The hyper-parameter α in our method is set to 6.0. The adversarial noise is crafted against the target classification model. The natural accuracy and robust accuracy are shown in Table. 2. The results show that our defense method achieve a great defense effect against general adversarial attacks. In addition, we use an expectation over transformation (EOT) attack. The adversarial samples generated by EOT are simultaneously adversarial over an entire distribution of transformations. The result in Table. 2 further indicates that our defense is effective to enhance the robustness of the deep learning model.

Defending against adaptive attacks. In addition to general adversarial attacks, a powerful *adaptive attack* strategy has been proposed to break defenses (Athalye et al., 2018a; Carlini & Wagner, 2017a). This strategy allows attackers access all information of defenses, including the architectures, model parameters and other proprietary procedures (*e.g.*, the pre-processing procedure). Thus, attackers can design targeted attacks.

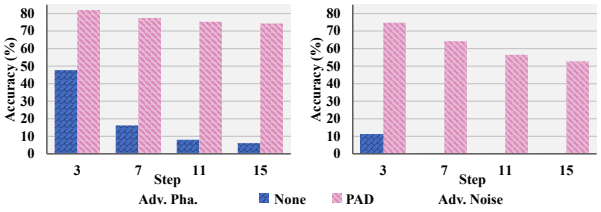

Figure 7. The robust accuracy against phase/pixel-level adversarial perturbations. *Adv. Pha.* shows the robust accuracy against phase-level perturbations and *Adv. Noise* presents the robust accuracy against pixel-level adversarial noise. *None* denotes the ordinarily-trained model. The attack is the same as the one in Figure. 3.

We design a white-box attack against the overall defense. The attack focuses on generating adversarial noise for the phase pattern. The optimization objective is given by

$$\max_{\delta} \ell_{ce}(h_\theta(\mathcal{F}^{-1}(\mathcal{F}_\phi(x+\delta), \mathcal{F}_\xi(\mathcal{T}(x+\delta, x_r)))), y), \quad (10)$$

where ℓ_{ce} is the cross-entropy loss. Similar to the operation in PGD, we compute the gradient of the loss and add it to the natural sample. Two classic pre-processing based defenses APE-G (Jin et al., 2019) and HGD (Liao et al., 2018) are used as baselines. They are trained together with the target model in an adversarial training manner. The adversarial noise are crafted by attacking the combination of the defense and the target model. In addition, we combine PGD-20 and a backward pass differentiable approximation (BPDA) attack to craft adversarial noise against the defense with pre-processing procedure. As shown in Table. 3, our defense presents better robust accuracy, which indicates the stability of our method.

Robustness on the phase pattern. We evaluate the effectiveness of defenses on the phase pattern. Similar to the experiment in Figure. 3, we compute the performance against the samples with only phase-level adversarial perturbations. For a fair comparison, the amplitude-based pre-processing procedure is not used in the inference stage. As shown in Figure. 7, our method achieves a large gain in accuracy against both the adversarial noise and phase-level perturbations. This result also reflects that enhancing phase-level robustness is beneficial for improving pixel-level robust accuracy. More comparisons are shown in Appendix. F. In addition, we perform adversarial training on both phase and amplitude patterns, which are presented in Appendix. G.

Table 4. Robust accuracy (percentage) of adversarially-trained models on *CIFAR-10*. PAT denotes the phase-level adversarial training, *i.e.*, the PAD with the pre-processing procedure removed.

Attack	None	PGD-20	AA	FAB
AT	83.34	47.23	44.45	46.28
APR	80.38	46.08	42.60	41.88
PAT	**83.63**	**51.50**	**47.61**	**48.57**

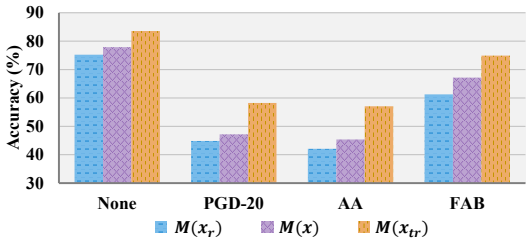

Figure 8. The impact of different reference samples on robust accuracy. $M(x)$/$M(x_r)$/$M(x_{tr})$ represents that using the amplitude spectrum of the natural/reference/transitional reference sample to combine with the adversarial phase spectrum, respectively. In the inference stage, $M(x)$ and $M(x_r)$ both use x_r to provide amplitude spectrum for the pre-processing procedure.

5.3. Ablation studies

Removing the pre-processing operation. In the inference stage, we remove the amplitude-based pre-processing operation and evaluate the inherent robustness of the model. We use AT and an amplitude-phase recombination (APR) method (Chen et al., 2021) as the baselines. As shown in Table. 4, the results show that focusing on the phase-level perturbations during the training process can improve the robust accuracy. This reflects the importance and criticality of the phase pattern to the robustness.

Transitional reference samples. To verify the effectiveness of the transitional reference sample x_{tr}, we replace x_{tr} with the natural sample x and the reference sample x_r, respectively. The former is to show the difference between Equation. 4 and Equation. 9, and the latter is to illustrate the benefits of constructing suitable reference samples. As shown in Figure. 6, x_{tr} can help reduce the obfuscated information in the recombined sample compared with x_r. The results in Figure. 8 indicate using x_{tr} can effectively improve both natural and robust accuracy.

The hyper-parameter. We explore the impact of the hyper-parameter α in Equation. 9 on defense effectiveness. α is set to 6.0 by default and adjusted to adjacent values in turn. The results in Figure. 9 show the positive effect of the similarity loss (lower accuracy when α=0), and present the stability of our method to the changes of α within a suitable range (the accuracy is almost unchanged when $\alpha \in [4, 8]$).

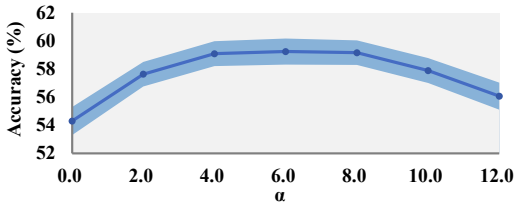

Figure 9. The impact of α on *CIFAR-10*. The light blue band presents the standard deviation of the accuracy against PGD-20.

6. Conclusion

The phase pattern in an image can reflect the structural semantics of objects and is a crucial concern in the human perception. Defending against adversarial noise from a phase perspective has not been deeply studied. In this paper, inspired by cognitive science, we explore the impact of the adversarial noise on the phase pattern and the robustness of the model against phase-level adversarial perturbations. Motivated by these observations, we propose a Phase-aware Adversarial Defense (PAD) method. This method designs a phase-level adversarial training to enhance the robustness on the phase pattern and constructs an amplitude-based pre-processing mechanism to mitigate the perturbation in the amplitude pattern. The empirical results show that our method can effectively defense the model against general and adaptive adversarial attacks. The limitation is that this work does not deeply consider more advanced style transfer techniques and amplitude-level measures. In future, we will design more powerful phase-level constraint and apply the phase-aware adversarial defense to other advanced robust learning frameworks (*e.g.*, causal-based methods) to further enhance the robustness of the model. Overall, our work is expected to provide a new defense strategy for the community of adversarial deep learning.

Acknowledgements

The authors greatly appreciate all reviewers. NNW was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0103202; in part by the National Natural Science Foundation of China under Grant U22A2096; in part by the Technology Innovation Leading Program of Shaanxi under Grant 2022QFY01-15; in part by Open Research Projects of the Zhejiang Laboratory under Grant 2021KG0AB01; in part by the Fundamental Research Funds for the Central Universities under Grant QTZX23042; in part by the Youth Innovation Team of Shaanxi University. XBG was supported by the National Natural Science Foundation of China under Grant 62036007. DWZ was supported in part by the Fundamental Research Funds for the Central Universities and in part by the Innovation Fund of Xidian University under Grant YJSJ23012.

References

Andrew, G. and Gao, J. Scalable training of l 1-regularized log-linear models. In *Proceedings of the 24th international conference on Machine learning*, pp. 33–40, 2007.

Athalye, A., Carlini, N., and Wagner, D. A. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, 2018a.

Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. Synthesizing robust adversarial examples. In *International conference on machine learning*, pp. 284–293. PMLR, 2018b.

Biederman, I. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94 (2):115, 1987.

Carlini, N. and Wagner, D. Magnet and" efficient defenses against adversarial attacks" are not robust to adversarial examples. *arXiv preprint arXiv:1711.08478*, 2017a.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 Ieee Symposium on Security and Privacy (sp)*, pp. 39–57. IEEE, 2017b.

Carlini, N., Tramer, F., Kolter, J. Z., et al. (certified!!) adversarial robustness for free! *arXiv preprint arXiv:2206.10550*, 2022.

Chen, G., Peng, P., Ma, L., Li, J., Du, L., and Tian, Y. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 458–467, 2021.

Clarysse, J., Hörmann, J., and Yang, F. Why adversarial training can hurt robust accuracy. *arXiv preprint arXiv:2203.02006*, 2022.

Concetta Morrone, M. and Burr, D. Feature detection in human vision: A phase-dependent energy model. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 235(1280):221–245, 1988.

Croce, F. and Hein, M. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning*, pp. 2196–2205. PMLR, 2020a.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning*, 2020b.

Ding, G. W., Lui, K. Y. C., Jin, X., Wang, L., and Huang, R. On the sensitivity of adversarial robustness to input data distributions. In *ICLR (Poster)*, 2019.

Freeman, J. and Simoncelli, E. P. Metamers of the ventral stream. *Nature neuroscience*, 14(9):1195–1201, 2011.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Ghiglia, D. C. and Pritt, M. D. Two-dimensional phase unwrapping: theory, algorithms, and software. *A Wiley Interscience Publication*, 1998.

Gladilin, E. and Eils, R. On the role of spatial phase and phase correlation in vision, illusion, and cognition. *Frontiers in Computational Neuroscience*, 9:45, 2015.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Guo, C., Frank, J. S., and Weinberger, K. Q. Low frequency adversarial perturbation. *arXiv preprint arXiv:1809.08758*, 2018a.

Guo, C., Rana, M., Cissé, M., and van der Maaten, L. Countering adversarial images using input transformations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018b.

Hansen, B. C. and Loschky, L. C. The contribution of amplitude and phase spectra-defined scene statistics to the masking of rapid scene categorization. *Journal of Vision*, 13(13):21–21, 2013.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Huang, X. and Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.

Jin, G., Shen, S., Zhang, D., Dai, F., and Zhang, Y. APE-GAN: adversarial perturbation elimination with GAN. In *International Conference on Acoustics, Speech and Signal Processing*, pp. 3842–3846, 2019.

Kaiming, H., Georgia, G., Piotr, D., and Ross, G. Mask r-cnn. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, PP:1–1, 2017.

Kovesi, P. Phase congruency: A low-level image invariant. *Psychological research*, 64(2):136–148, 2000.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Landau, B., Smith, L., and Jones, S. Object shape, object function, and object name. *Journal of memory and language*, 38(1):1–27, 1998.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, Q., Guo, Y., Zuo, W., and Chen, H. Squeeze training for adversarial robustness. In *The Eleventh International Conference on Learning Representations*, 2023.

Li, S., Xia, X., Ge, S., and Liu, T. Selective-supervised contrastive learning with noisy labels. In *CVPR*, pp. 316–325, 2022.

Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., and Zhu, J. Defense against adversarial attacks using high-level representation guided denoiser. In *Conference on Computer Vision and Pattern Recognition*, pp. 1778–1787, 2018.

Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S. N. R., Schoenebeck, G., Song, D., Houle, M. E., and Bailey, J. Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations*, 2018.

Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., and Lu, F. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110:107332, 2021.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations*, 2018.

Morrone, M. C. and Owens, R. A. Feature detection from local energy. *Pattern recognition letters*, 6(5):303–313, 1987.

Morrone, M. C., Ross, J., Burr, D. C., and Owens, R. Mach bands are phase dependent. *Nature*, 324(6094):250–253, 1986.

Naseer, M., Khan, S., Hayat, M., Khan, F. S., and Porikli, F. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 262–271, 2020.

Olivier, R., Raj, B., and Shah, M. High-frequency adversarial defense for speech and audio. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2995–2999. IEEE, 2021.

Oppenheim, A. V. and Lim, J. S. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981.

Pang, T., Yang, X., Dong, Y., Su, H., and Zhu, J. Bag of tricks for adversarial training. *arXiv preprint arXiv:2010.00467*, 2020.

Pollen, D. A. and Ronner, S. F. Phase relationships between adjacent simple cells in the visual cortex. *Science*, 212 (4501):1409–1411, 1981.

Pollen, D. A. and Ronner, S. F. Visual cortical neurons as localized spatial frequency filters. *IEEE Transactions on Systems, Man, and Cybernetics*, (5):907–916, 1983.

Rice, L., Wong, E., and Kolter, Z. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020.

Samuelson, L. K. and Smith, L. B. They call it like they see it: Spontaneous naming and attention to shape. *Developmental Science*, 8(2):182–198, 2005.

Schmidt, F., Kleis, J., Morgenstern, Y., and Fleming, R. W. The role of semantics in the perceptual organization of shape. *Scientific Reports*, 10(1):1–19, 2020.

Sharma, Y., Ding, G. W., and Brubaker, M. On the effectiveness of low frequency perturbations. *arXiv preprint arXiv:1903.00073*, 2019.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations*, 2015.

Song, Z. and Deng, Z. An adversarial examples defense method based on image low-frequency information. In *International Conference on Artificial Intelligence and Security*, pp. 204–213. Springer, 2021.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Tolhurst, D. J., Tadmor, Y., and Chao, T. Amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, 12(2):229–232, 1992.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Wang, H., Wu, X., Huang, Z., and Xing, E. P. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8684–8694, 2020.

Wang, H., Zhang, A., Zheng, S., Shi, X., Li, M., and Wang, Z. Removing batch normalization boosts adversarial training. In *International Conference on Machine Learning*, pp. 23433–23445. PMLR, 2022.

Wang, X. and He, K. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1924–1933, 2021.

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., and Gu, Q. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2019.

Wu, D., Xia, S.-T., and Wang, Y. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33, 2020.

Xia, X., Liu, T., Han, B., Wang, N., Gong, M., Liu, H., Niu, G., Tao, D., and Sugiyama, M. Part-dependent label noise: Towards instance-dependent label noise. *Advances in Neural Information Processing Systems*, 33, 2020.

Xia, X., Liu, J., Yu, J., Shen, X., Han, B., and Liu, T. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *ICLR*, 2023.

Xiao, C., Zhu, J., Li, B., He, W., Liu, M., and Song, D. Spatially transformed adversarial examples. In *6th International Conference on Learning Representations*, 2018.

Yin, D., Gontijo Lopes, R., Shlens, J., Cubuk, E. D., and Gilmer, J. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019.

Yu, C., Han, B., Shen, L., Yu, J., Gong, C., Gong, M., and Liu, T. Understanding robust overfitting of adversarial training and beyond. In *International Conference on Machine Learning*, pp. 25595–25610. PMLR, 2022a.

Yu, C., Zhou, D., Shen, L., Yu, J., Han, B., Gong, M., Wang, N., and Liu, T. Strength-adaptive adversarial training. *arXiv preprint arXiv:2210.01288*, 2022b.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In Wilson, R. C., Hancock, E. R., and Smith, W. A. P. (eds.), *Proceedings of the British Machine Vision Conference 2016*, 2016.

Zhang, F., Jiang, W., Autrusseau, F., and Lin, W. Exploring v1 by modeling the perceptual quality of images. *Journal of Vision*, 14(1):26–26, 2014.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482. PMLR, 2019.

Zhang, L., Zhang, L., Mou, X., and Zhang, D. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

Zhang, Y., Zhang, G., Khanduri, P., Hong, M., Chang, S., and Liu, S. Revisiting and advancing fast adversarial training through the lens of bi-level optimization. In *International Conference on Machine Learning*, pp. 26693–26712. PMLR, 2022.

Zhou, D., Liu, T., Han, B., Wang, N., Peng, C., and Gao, X. Towards defending against adversarial examples via attack-invariant features. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 12835–12845, 2021a.

Zhou, D., Wang, N., Gao, X., Han, B., Wang, X., Zhan, Y., and Liu, T. Improving adversarial robustness via mutual information estimation. In *International Conference on Machine Learning*, pp. 27338–27352. PMLR, 2022.

Zhou, Y., Hu, X., Han, J., Wang, L., and Duan, S. High frequency patterns play a key role in the generation of adversarial examples. *Neurocomputing*, 459:131–141, 2021b.

Zibar, D., de Carvalho, L. H. H., Piels, M., Doberstein, A., Diniz, J., Nebendahl, B., Franciscangelis, C., Estaran, J., Haisch, H., Gonzalez, N. G., et al. Application of machine learning techniques for amplitude and phase noise characterization. *Journal of Lightwave Technology*, 33(7):1333–1343, 2015.

A. Related work

Adversarial attacks. Adversarial noise is an imperceptible but seriously misleading noise crafted by adversarial attack algorithms (Szegedy et al., 2014; Goodfellow et al., 2015). The sample with adversarial noise added is called the adversarial sample. Existing studies have proposed a variety of adversarial attack strategies. For example, some works proposed to craft adversarial noise following the gradient direction of loss functions, such as projected gradient descent (PGD) attack (the strongest first-order attack) (Madry et al., 2018), powerful autoattack (AA) (Croce & Hein, 2020b) and variance tuning momentum iterative fast gradient sign method (VMI-FGSM) (Wang & He, 2021). These adversarial noise is bounded by a small norm-ball $\|\cdot\|_p \leq \epsilon$, so that their adversarial samples are visually similar to natural samples for humans. Some works designed optimization-based strategies, such as fast adaptive boundary (FAB) attack (Croce & Hein, 2020a) and Carlini&Wagner (C&W) attack (Carlini & Wagner, 2017b). These works aimed to minimize the perturbation size while ensuring that the predictions of models are wrong. Moreover, some attacks, such as spatial transform attack (STA) (Xiao et al., 2018) aims to perturb the spatial structure information of the objective.

In addition to above attack algorithms, we examine the defense by using a backward pass differentiable approximation (BPDA) attack (Athalye et al., 2018a) and a expectation over transformation (EOT) attack (Athalye et al., 2018b). In addition, we design specific adaptive attacks against the phase pattern to comprehensively verify the effectiveness of the proposed method.

Adversarial defenses. The vulnerability of deep learning models to adversarial noise promotes the development of adversarial defenses. A representative defense strategy is devoted to enhancing the adversarial robustness of models in an adversarial training manner (Madry et al., 2018; Ding et al., 2019). Methods based on this strategy utilize adversarial noise to augment training data and train the model via a min-max optimization formulation. Some works modified or reformulated the regularization term to perform more effective adversarial training (Zhang et al., 2019; Wang et al., 2019). The work in Ilyas et al. (2019) constructed robust training set to guide the model to learn robust features and achieve robust performances. The works in Yin et al. (2019); Wang et al. (2020); Olivier et al. (2021); Zhou et al. (2021b) analyzed adversarial noise in the high-frequency component and devised targeted methods to enhance robustness. The studies in (Guo et al., 2018a; Sharma et al., 2019; Song & Deng, 2021) developed adversarial noise against low-frequency information, which provokes defenses on the low-frequency component.

Differently, our work designs adversarial training from a phase perspective. The phase pattern is closely consistent with structural information. Note that although high-frequency components can exhibit some boundary information, they essentially reflect the parts with rapidly changing pixel intensities, which cannot accurately represent structural semantics (see Figure. 10) and are not perceivable to humans (Wang et al., 2020; Ilyas et al., 2019). Although the work in (Chen et al., 2021) discussed the influence of phase on the generalization behavior of convolutional neural networks, the impact of adversarial noise on the phase pattern and the defense against phase-level perturbations have not been well studied. Moreover, some works use data selection (Xia et al., 2023) or dependence relations (Li et al., 2022; Xia et al., 2020) to explore the negative effects on target models from the intrinsic components of noisy data, which have similar motivations to our work.

In addition, the pre-processing based defense strategy has also been extensively studied. This strategy typically aims to remove adversarial noise by learning denoising maps or feature-squeezing functions. For example, the works in Liao et al. (2018); Naseer et al. (2020) learned a mapping from adversarial data to natural data via natural-adversarial data pairs, to remove adversarial noise in the inference stage. The work in Guo et al. (2018b) utilized the learned feature-squeezing function to reduce adversarial and redundant information. However, the denoising models themselves are likely to be corrupted by adversarial noise and lose their functionality, which brings new threats. Differently, the pre-processing mechanism in our method mainly involves the replacement of the amplitude spectrum, which is based on the rigorous Fourier transform.

B. Qualitative study on phase-level adversarial perturbations

We concretize the phase pattern of the image by performing the inverse Fourier transform on only the phase spectrum (replacing the amplitude spectrum with a constant matrix). As shown in Figure. 11, adversarial noise perturb or eliminate some structural semantics of the objective. For example, the snake scales, the contour of the bird's head, the bucket and the dog's nose are significantly perturbed. We note that these features are often not the core information for identifying birds or snakes, as we humans can still clearly recognize them via the beak, claws or overall contour features. They may be

predictable but easily disturbed phase features.

C. Discrete Fourier transform

Since the work in this paper is mainly concerned with adversarial attacks and defenses on images, we utilize the discrete Fourier transforms (DFT). We denote x in the frequency domain by F_x, the amplitude spectrum of x by ξ_x and the phase spectrum of x by ϕ_x. For an image x of size $H \times W$, its mapping in frequency domain is as follows:

$$F_x(u,v) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x(h,w) e^{-i2\pi(\frac{uh}{H}+\frac{vw}{W})}, \tag{11}$$

where (h,w) denotes the pixel coordinates, (u,v) denotes the coordinates of spectrum values, $u = 0, 1, 2, \ldots, H-1$ and $v = 0, 1, 2, \ldots, W-1$. The amplitude and phase of F_x are as follows:

$$\begin{aligned} \xi_x(u,v) &= \sqrt{\mathrm{Re}(F_x(u,v))^2 + \mathrm{Im}(F_x(u,v))^2}, \\ \phi_x(u,v) &= \arctan\left[\frac{\mathrm{Im}(F_x(u,v))}{\mathrm{Re}(F_x(u,v))}\right], \end{aligned} \tag{12}$$

where Re and Im denote the real and imaginary signals of F_x. The image x can be recovered from F_x via the inverse Discrete Fourier transform (IDFT):

$$x(h,w) = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} F_x(u,v) e^{i2\pi(\frac{uh}{H}+\frac{vw}{W})}. \tag{13}$$

For a more intuitive presentation, we reformulate Equation. 11 and 13 in matrix form as:

$$F_x = \mathcal{F}(x) = \xi_x \otimes e^{i\cdot\phi_x}, \tag{14}$$

$$x = \mathcal{F}^{-1}(F_x) = \mathcal{F}^{-1}(\xi_x \otimes e^{i\cdot\phi_x}), \tag{15}$$

where $\otimes$ denotes the element-wise multiplication of two matrices, $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier transform and the inverse Fourier transform, respectively.

D. Experiment setup

Dataset. In this paper, we use two classic datasets *CIFAR-10* (Krizhevsky et al., 2009) and *Mini-ImageNet* (Vinyals et al., 2016) to evaluate the effectiveness of our method. *CIFAR-10* has 10 classes of images with a resolution of 32×32. *Mini-ImageNet* has 100 classes of images with a resolution of 84×84. They both have 50,000 training images and 10,000 test images. All images are normalized into [0,1], and performed simple data augmentations in the training process, including random crop and random horizontal flip. For the target model, we mainly utilize ResNet-18 (RN) (He et al., 2016) for *CIFAR-10* and Wide ResNet-28-10 (WRN) for *Mini-ImageNet* (Zagoruyko & Komodakis, 2016).

Attack settings. We utilize seven types of adversarial attacks to evaluate the performances of defenses. They are L_∞-norm PGD (Madry et al., 2018), L_∞-norm AA (Croce & Hein, 2020b), L_∞-norm VMIFGSM(Wang & He, 2021), L_∞-norm EOT-PGD (Athalye et al., 2018b), L_∞-norm FAB(Croce & Hein, 2020a), L_2-norm CW (Carlini & Wagner, 2017b) and STA (Xiao et al., 2018). Among them, the AA attack algorithm integrates three non-target attacks and a target attack. Other attack algorithms belong to non-target attacks. The perturbations crafted by STA mainly make small modifications to the contour of the objective. The iteration numbers for PGD, EOT-PGD, VMIFGSM and FAB are set to 20, and those for STA and CW are 10 and 200, respectively. For *CIFAR-10* and *Mini-ImageNet*, the perturbation budget ϵ for L_∞-norm attacks is set to $8/255$.

Defense settings. We use three representative adversarial training methods as the baselines: standard adversarial training (AT) (Madry et al., 2018), optimized adversarial training methods TRADES (Zhang et al., 2019) and MART (Wang et al., 2019). In addition, we use two great pre-processing based defense as additional baselines. For all defense methods, we use the L_∞-norm non-target PGD-10 with random start and step size $\epsilon/4$ to craft adversarial noise in the training stage. The perturbation budget ϵ is set to $8/255$ for both *CIFAR-10* and *Tiny-ImageNet*. All the defense models are trained using SGD

(Andrew & Gao, 2007) with momentum 0.9. For adversarial training, the initial learning rate is set to $2e^{-1}$ corresponding to the batch size 256 according to work in Pang et al. (2020). The weight decay is 2×10^{-4} for *CIFAR-10*, and is 5×10^{-4} for *Mini-ImageNet*. We set $\lambda = 6$ for TRADES and MART. The epoch number is set to 91 by using the ear-stopping strategy. For pre-processing based methods, we use the settings from their original papers.

E. Examples of obfuscated information

The phase-level adversarial samples (*i.e.*, recombined samples) are shown in Figure. 12. The obfuscated information in the recombined samples using transitional reference sample is significantly less than that in the recombined samples using general reference samples. The objective can be clearly seen in the former.

F. Robustness on the phase pattern

We investigate the performances against the samples with only phase-level adversarial perturbations. As shown in Figure. 13, compared with ordinary training mechanism and standard adversarial training (AT), our method achieves more gains in accuracy against both the adversarial noise and phase-level perturbations.

G. Adversarial training on both phase and amplitude patterns

In this work, we mainly perform phase-level adversarial training (PAT). Here, we further consider adversarial training on phase and amplitude patterns, *i.e.*, phase-amplitude-level adversarial training (PAAT). To obtain the clean phase pattern, we utilize the corresponding natural sample as the reference for the replacement of the phase spectrum. The classification loss is formulated as:

$$\begin{aligned}\mathcal{L}_c(x, x', y) = -\frac{1}{N}\sum_{i=1}^{N}[\boldsymbol{y_i} \cdot \log(h_\theta(\mathcal{F}^{-1}(\phi_{x_i'}, \xi_{x_{tr_i}}))) \\ + \boldsymbol{y_i} \cdot \log(h_\theta(\mathcal{F}^{-1}(\phi_{x_i}, \xi_{x_i'})))],\end{aligned} \tag{16}$$

where h_θ denotes the target model, $\boldsymbol{y}$ denotes the one-hot label, x denotes the natural sample, x' denotes the adversarial sample, x_{tr} denotes the transitional reference sample, ϕ_x denotes the phase spectrum of x and ξ_x denotes the amplitude spectrum of x. The optimization objective is the same as Eq.9 in the main text. The experimental results are shown in Table. 5. Adding amplitude-level adversarial training does not bring significant improvement or even cause some declines. This indirectly shows that the phase pattern may be a more critical role for the adversarial robustness.

Table 5. Robust accuracy (percentage) on *CIFAR-10*. The pre-processing procedure is not used.

Attack	None	PGD-20	AA	FAB
AT	83.34	47.23	44.45	46.28
PAT	83.63	51.50	47.61	48.57
PAAT	83.41	50.06	46.27	47.19

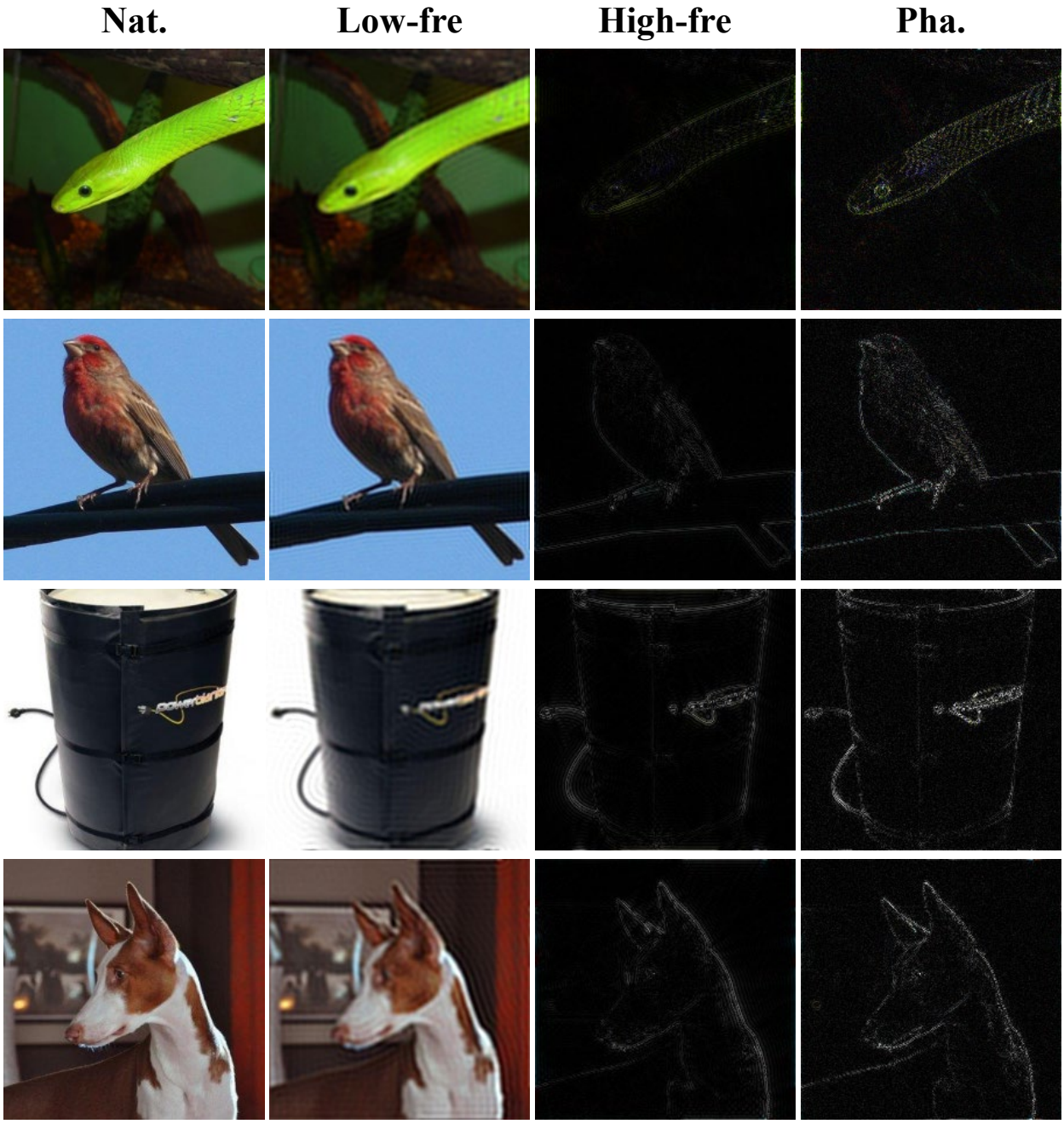

Figure 10. The low-frequency components, high-frequency components and phase patterns of images. *Nat.*, *Low-fre.*, *High-fre.* and *Pha.* denote the natural sample, low-frequency component, high-frequency component and the phase pattern, respectively.

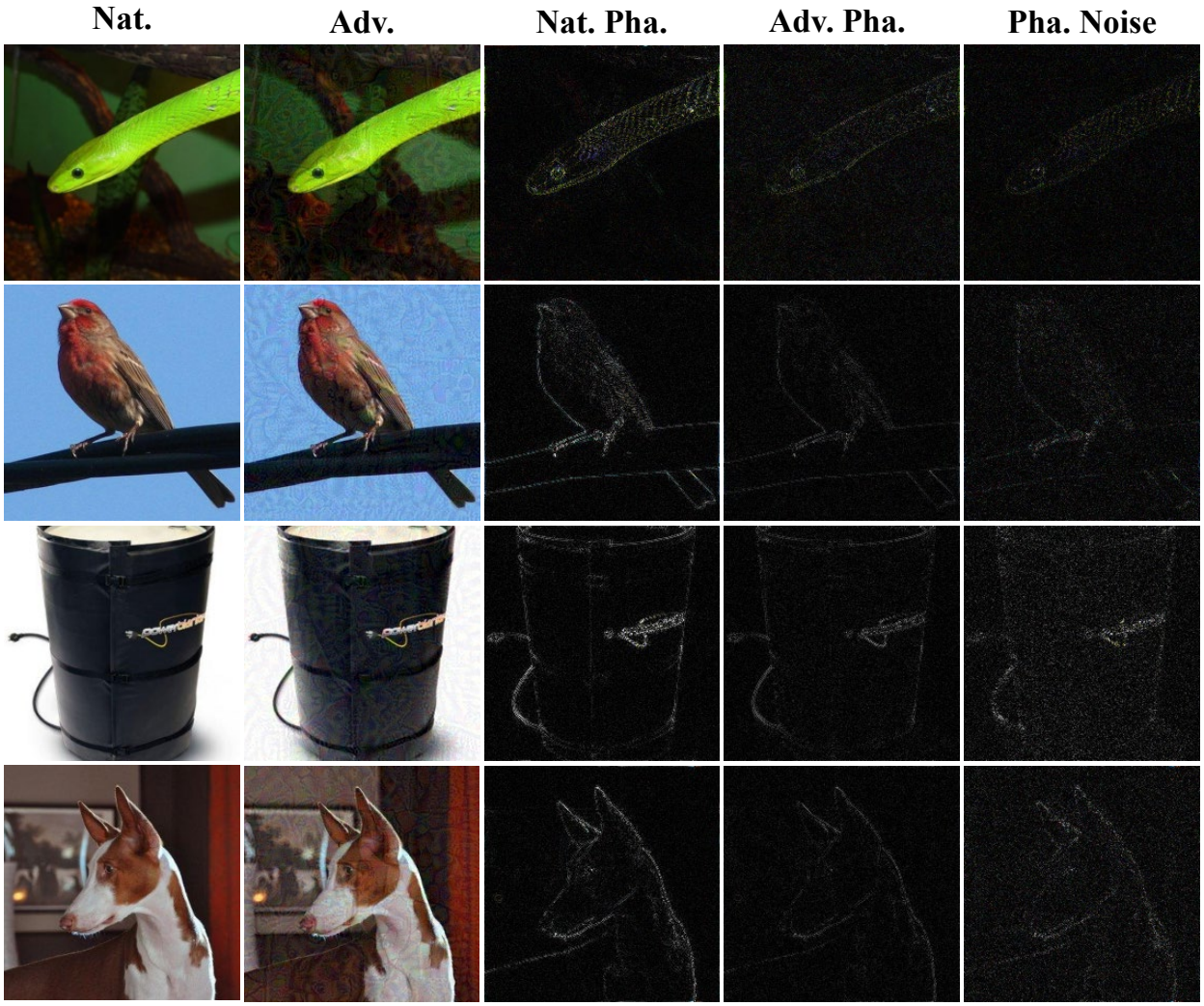

Figure 11. The examples of the natural phase patterns and the perturbed phase patterns. *Nat.*, *Adv.*, *Nat. Pha.*, *Adv. Pha.* and *Pha. Noise.* denote the natural sample, adversarial sample, natural phase pattern, adversarial phase pattern and phase-level adversarial noise, respectively.

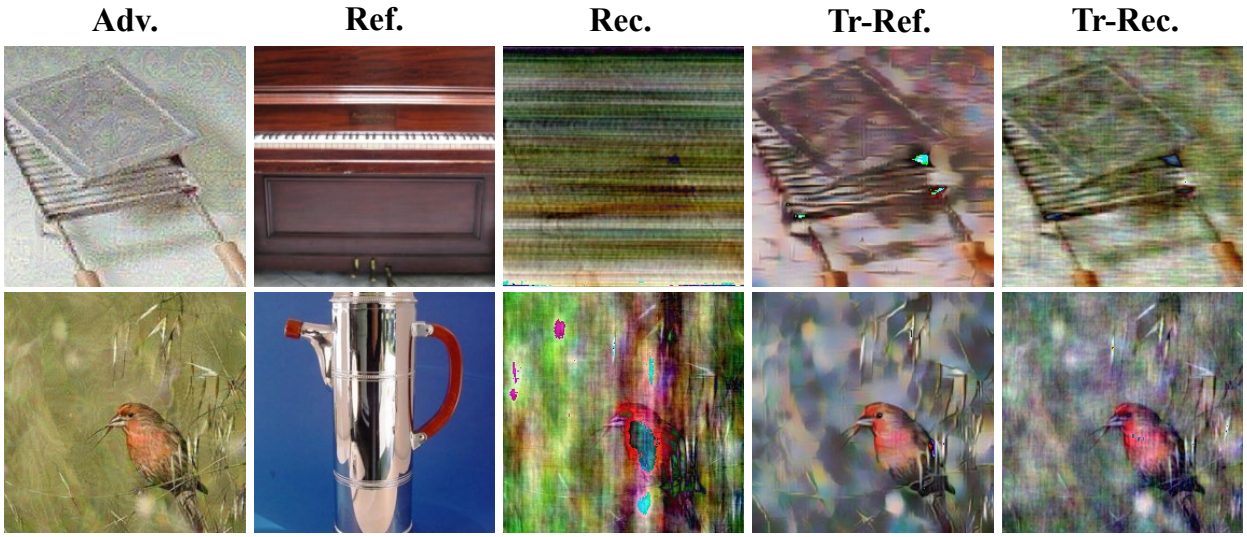

Figure 12. The phase-level adversarial samples (recombined samples) generated using different reference samples. *Adv.*, *Ref.*, *Rec.*, *Tr-Ref.* and *Tr-Rec.* denote the adversarial sample, original reference sample, the recombined sample with *Ref.*, the transitional reference sample and the recombined sample with *Tr-Ref.*. The obfuscated information in *Tr-Rec.* is significantly less than that in *Rec.* and the objective can be clearly seen in *Tr-Rec.*.

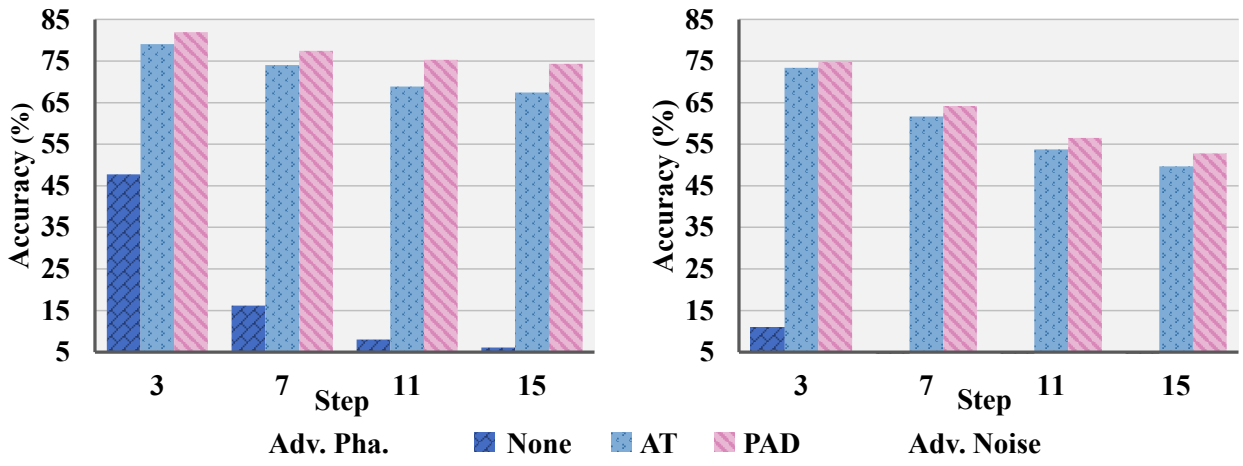

Figure 13. The robust accuracy against phase/pixel-level adversarial perturbations. *Adv. Pha.* shows the robust accuracy against phase-level perturbations and *Adv. Noise* presents the robust accuracy against pixel-level adversarial noise. *None* denotes the ordinarily-trained model. The attack PGD with perturbation budget 8/255 and step size 1/255.