Towards Compositional Adversarial Robustness: Generalizing Adversarial Training to Composite Semantic Perturbations

Lei Hsiung[1,4], Yun-Yun Tsai[2], Pin-Yu Chen[3], Tsung-Yi Ho[1,4]
[1]National Tsing Hua University, [2]Columbia University, [3]IBM Research,
[4]The Chinese University of Hong Kong

https://hsiung.cc/CARBEN/

Abstract

*Model robustness against adversarial examples of single perturbation type such as the ℓ_p-norm has been widely studied, yet its generalization to more realistic scenarios involving multiple semantic perturbations and their composition remains largely unexplored. In this paper, we first propose a novel method for generating composite adversarial examples. Our method can find the optimal attack composition by utilizing component-wise projected gradient descent and automatic attack-order scheduling. We then propose **generalized adversarial training** (**GAT**) to extend model robustness from ℓ_p-ball to composite semantic perturbations, such as the combination of Hue, Saturation, Brightness, Contrast, and Rotation. Results obtained using ImageNet and CIFAR-10 datasets indicate that GAT can be robust not only to all the tested types of a single attack, but also to any combination of such attacks. GAT also outperforms baseline ℓ_∞-norm bounded adversarial training approaches by a significant margin.*

1. Introduction

Deep neural networks have shown remarkable success in a wide variety of machine learning (ML) applications, ranging from biometric authentication (e.g., facial image recognition), medical diagnosis (e.g., CT lung cancer detection) to autonomous driving systems (traffic sign classification), etc. However, while these models can achieve outstanding performance on benign data points, recent research has shown that state-of-the-art models can be easily fooled by malicious data points crafted intentionally with adversarial perturbations [37].

To date, the most effective defense mechanism is to incorporate adversarial examples during model training, known as adversarial training (AT) [21, 48]. Nonetheless, current adversarial training approaches primarily only consider a single perturbation type (or threat model) quantified in a specific distance metric (e.g., ℓ_p-ball). In this regard, the lack of exploration of the compositional adversarial robustness against a combination of several threat models could lead to impractical conclusions and undesirable bias in robustness evaluation. For example, a model that is robust to perturbations within ℓ_p-ball does not imply it can simultaneously be robust to other realistic semantic perturbations (e.g., hue, saturation, rotation, brightness, and contrast).

To tackle this issue, in this paper, we propose **generalized adversarial training (GAT)**, which can harden against a wide range of threat models, from single ℓ_∞-norm or semantic perturbation to a combination of them. Notably, extending standard adversarial training to composite adversarial perturbations is a challenging and non-trivial task, as each perturbation type is sequentially applied, and thus the attack order will affect the effectiveness of the composite adversarial example. To bridge this gap, we propose an efficient attack order scheduling algorithm to learn the optimal ordering of various perturbation types, which will then be incorporated into the GAT framework.

Different from existing works, this paper aims to address the following fundamental questions: (a) How to generalize adversarial training from a single threat model to multiple? (b) How to optimize the perturbation order from a set of semantic and ℓ_p-norm perturbations? (c) Can GAT outperform other adversarial training baselines against composite perturbations?

Our main contributions in this paper provide affirmative answers to the questions:

1. We propose composite adversarial attack (CAA), a novel and unified approach to generate adversarial examples across from multiple perturbation types with attack-order-scheduling, including semantic perturbations (*Hue, Saturation, Contrast, Brightness and Contrast*) and ℓ_p-norm space. To the best of our knowledge, this paper is the first work that leverages a scheduling algorithm for finding the optimal attack order in composite adversarial attacks.

2. Building upon our composite adversarial attack framework, we propose generalized adversarial training (**GAT**) toward achieving compositional adversarial robustness,

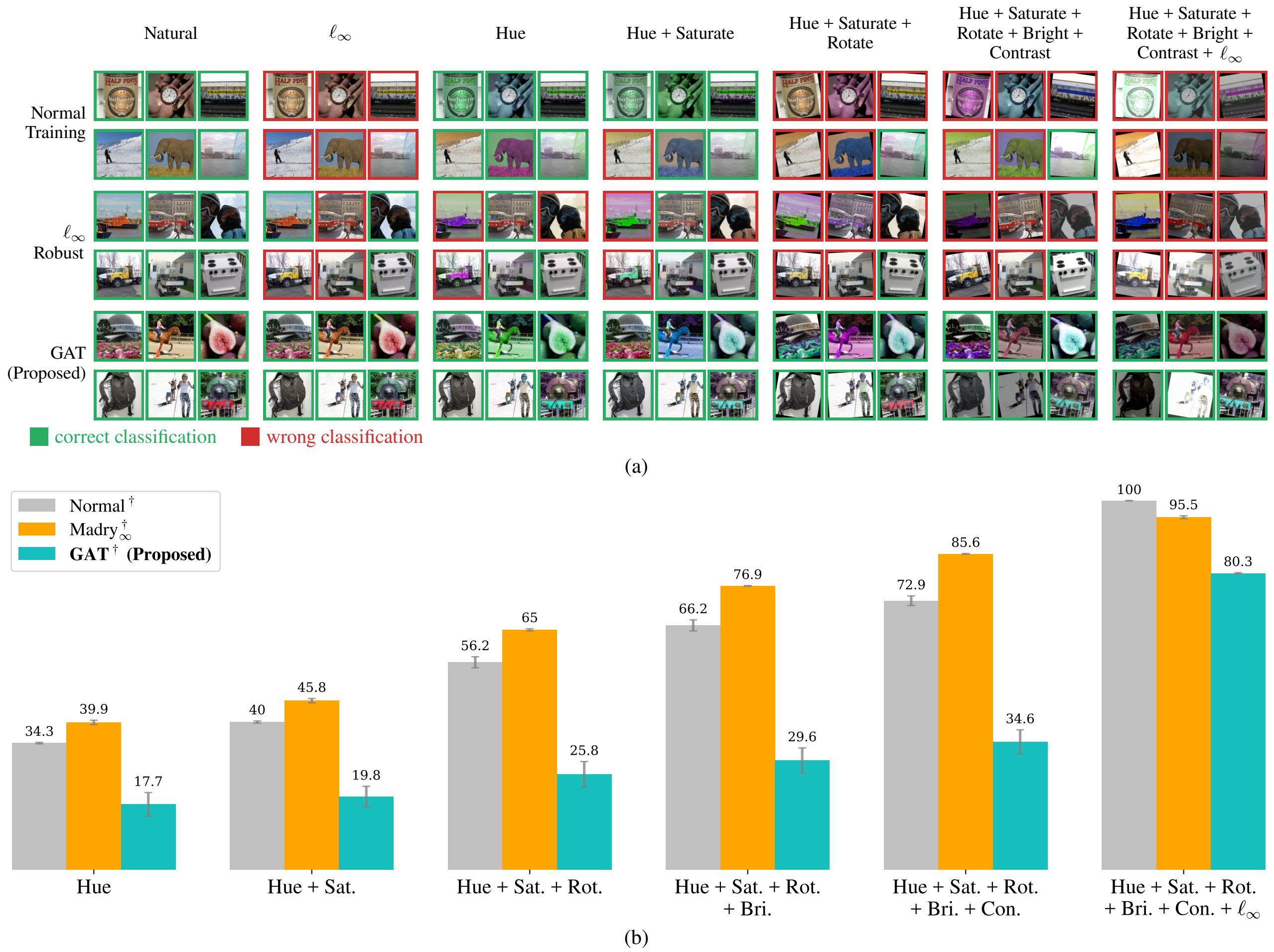


Figure 1. (a) **Qualitative study:** illustration of some perturbed examples generated by different attack combinations and their predictions by different ResNet50 models [10] on ImageNet, including standard training, Madry's ℓ_∞ robust training [21] and our proposed **GAT**. The results show that our proposed GAT can maintain robust accuracy under a variety of composite adversarial attacks, even with the increasing number of attacks. (b) **Quantitative study:** the attack success rate (ASR, %) of the above-mentioned models under multiple composite attacks (a higher ASR means less robust) on all correctly classified test samples for each model. The corresponding robust accuracy (RA) is listed in Table 3.

which enables the training of neural networks robust to composite adversarial attacks.

3. For the attack part, our proposed composite adversarial attack exhibits a high attack success rate (ASR) against standard or ℓ_∞-norm robust models. Moreover, our method with learned attack order significantly outperforms random attack ordering, giving an average 9% and 7% increase in ASR on CIFAR-10 and ImageNet.

4. For the defense part, comparing our GAT to other adversarial training baselines [20, 21, 42, 44, 48, 49], the results show the robust accuracy of GAT outperforms them by average 30% $\sim$ 60% on semantic attacks and 15% $\sim$ 22% on full attacks.

To further motivate the effectiveness of our proposed GAT framework, Fig. 1 compares the performance of different models under selected attacks, ranging from a single threat to composite threats. The models include standard training, ℓ_∞-robust training, and our proposed GAT. The results show the limitation of ℓ_∞-robust model [21], which is robust against the same-type adversarial attack, but becomes fragile against semantic adversarial attacks and their composition. Our proposed GAT addresses this limitation by providing a novel training approach that is robust to any combination of multiple and adversarial threats.

2. Related Work

Adversarial Semantic Perturbations. Adversarial machine learning research has largely focused on generating examples that can deceive models into making incorrect predictions [2]. One widely studied class of attacks involves ℓ_p-norm adversarial perturbations [3, 4, 6, 9]. However, natural transformations such as changes in geometry, color, and brightness can also cause adversarial vulnerabilities, leading to what are known as semantic perturbations [1, 11, 13, 14, 29, 32, 40, 41]. Unlike ℓ_p-norm perturbations, semantic perturbations often result in adversarial examples that look natural and are semantically similar to the original image but have significant differences in the ℓ_p-norm perspective.

Hosseini and Poovendran [11] demonstrated that randomly shifting the Hue and Saturation components in the Hue-Saturation-Value (HSV) color space of images can significantly reduce the accuracy of a neural network by up to 88%. Bhattad et al. [1] proposed similar attacks, including colorization and texture transfer attacks, which can perturb a grayscale image with natural colorization or infuse the texture of one image into another. Prior work on geometric transformations has targeted rotation transformations. For example, Xiao et al. [45] used coordinate-wise optimization at each pixel, which can be computationally expensive. Engstrom et al. [8] proposed a simple way of parametrizing a set of tunable parameters for spatial transformations. Dunn et al. [7] exploited context-sensitive changes to features from the input and perturbed images with the corresponding feature map interpolation. Mohapatra et al. [26] studied certified robustness against semantic perturbations, but they did not discuss adversarial training.

Composite Adversarial Perturbations. Previous literature has inspired researchers to explore different metrics [20, 43] and the combination of various adversarial threats [1, 12, 19, 46] to harden adversarial examples. These works have expanded the perturbation space of an image and have successfully increased the misclassification rate of neural networks. For instance, Laidlaw and Feizi [19] propose the ReColorAdv attack, which combines multi-functional threats to perturb every input pixel and also includes additional ℓ_p-norm threat. Mao et al. [23] utilized genetic algorithms to search for the best combination of multiple attacks, which were found to be stronger than a single attack. However, their approach only considered searching the order of attack combination in specific norm spaces (i.e., ℓ_2, ℓ_∞, and corruption semantic space) and could not handle all attacks simultaneously. On the other hand, Yuan et al. [46] have incorporated different image transformation operations to improve the transferability of adversarial examples.

Regarding measuring model robustness, Kang et al. [14] proposed utilizing ensemble unforeseen attacks from broader threat models, including JPEG, Fog, Snow, Gabor, etc. They consider the worst-case scenario over all attacks and attempt to improve model performance against these unforeseen adversarial threats. Prior works have shown that combining different types of adversarial threats can result in more robust adversarial examples. Our work builds upon these ideas and proposes a method for scheduling multiple attack types to generate composite adversarial perturbations.

Adversarial Training (AT). AT is a widely adopted method for improving model robustness against adversarial attacks [18, 21, 48, 50]. One of the pioneering works in this field is by Madry et al. [21], who proposed to minimize the worst-case loss in a region around the input. Zhang et al. [48] further improved AT by considering both natural and adversarial inputs in computing the loss, along with a parameter β to define the ratio of them, resulting in a smoother robust decision boundary. Laidlaw et al. [20] expanded adversarial attacks from single to multiple threat models by using neural perceptual distance measurement to generalize adversarial training with perceptual adversarial examples. Recently, Mao et al. [24] proposed to combine robust components as building blocks of vision transformers, leading to a state-of-the-art robust vision transformer. AT with adversarial transformations is also done in [8, 36].

While most previous works on AT have focused on improving model robustness against a single threat model, as shown in Fig. 1, a model that is robust against ℓ_∞-norm perturbations may still have low robustness against composite semantic attacks or other ℓ_q threats ($p \neq q$) [33]. This has led researchers to consider multiple-norm adversarial training [22, 38, 39], which yields models that are simultaneously robust against multiple ℓ_p-norm attacks. Tramer et al. [38] have considered alternately optimizing perturbation types given a fixed attack order, but the search for the strongest possible attack order is left out of their discussion. Also, the considered perturbations are simultaneously added to the same data sample rather than sequentially.

In contrast to prior arts, this paper offers a novel approach to improving model robustness against multiple adversarial threats. Our attack takes into account the efficient attack order scheduling and extends beyond the ℓ_p-norm attacks by incorporating various semantic perturbations, which can result in more robust adversarial examples. By incorporating the composite adversarial examples, our defense mechanism can significantly improve the robustness of the model. Overall, our work represents an important step towards creating more robust deep learning models that can defend against a wide range of adversarial attacks.

3. Methodologies (CAA & GAT)

In this section, we first propose the composite adversarial attack (CAA) framework (Fig. 2), and elucidate the details of our attack order scheduling algorithm. We then adopt the CAA into adversarial training, which is called generalized adversarial training (GAT).

3.1. Composite Attack Formulation

Order Scheduling. Let $\mathcal{F} : \mathcal{X} \to \mathbb{R}^d$ be an image classifier that takes image $x \in \mathcal{X}$ as input and generates a d-dimensional prediction scores (e.g., softmax outputs) for d classes, and let $\Omega = \{A_1, \ldots, A_n\}$ denote an at-

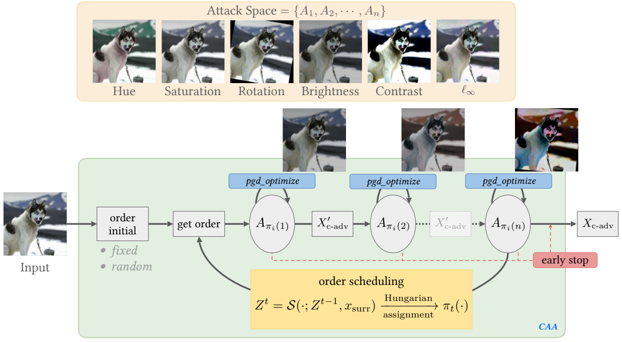


Figure 2. A pipeline of the proposed *composite adversarial attack* method with the ability to dynamically optimize the attack order and harden adversarial examples.

tack set that contains n attack types. For each attack A_k, we define a corresponding perturbation interval (boundary) $\epsilon_k = [\alpha_k, \beta_k]$ to govern the attack power of A_k. We then denote the corresponding perturbation intervals of Ω as $E = \{\epsilon_k | k \in \{1, \dots, n\}\}$.

In CAA, we optimize not only the power of each attack component in Ω, but also the attack order applied to the image x. That is, consider $\mathcal{I}_n = \{i\}_{i=1}^n$, we can use an assignment function $\pi_i : \mathcal{I}_n \to \mathcal{I}_n$ to determine the attack order to be used under the i-th schedule. As shown in Fig. 2, after i-th scheduling, a composite adversarial example $x_{\text{c-adv}}$ can be formulated as:

$$x_{\text{c-adv}} = A_{\pi_i(n)}(A_{\pi_i(n-1)}(\cdots A_{\pi_i(1)}(x))).$$

Noted that input x would be perturbed in the order of: $A_{\pi_i(1)} \to A_{\pi_i(2)} \to \cdots \to A_{\pi_i(n)}$. For each attack operation $A_k \in \Omega$, an input x would be transformed to a perturbed sample with a specific perturbation level δ_k, where $\delta_k \in \epsilon_k$ would be optimized via projected gradient descent, maximizing the classification error (e.g., cross-entropy loss $\mathcal{L}$). Therefore, the operation of $A_k(x; \delta_k)$ could be expressed as optimizing δ_k, that is:

$$\arg\max_{\delta_k \in \epsilon_k} \mathcal{L}(\mathcal{F}(A_k(x; \delta_k)), y), \tag{1}$$

where y is the ground-truth label of x. We named it component-wise PGD (Comp-PGD) and will explain more details in Sec. 3.2.

Since the assignment function $\pi_i(\cdot)$ is essentially a permutation matrix (or Birkhoff polytope), we can optimize it by treating it as a (relaxed) *scheduling matrix* Z^i, where $Z^i = \left[\mathbf{z}_1, \dots, \mathbf{z}_n\right]^\top$ is also a doubly stochastic matrix, i.e. $\mathbf{z}_j \in \mathbb{R}^n$, $\sum_i z_{ij} = \sum_j z_{ij} = 1, \forall\, i, j \in \{1, \dots, n\}$. Furthermore, we can utilize the Hungarian algorithm [16, 27] to obtain an optimal attack order assignment.

In sum, we formalize CAA's attack order auto-scheduling as a constrained optimization problem, where the attack order having maximum classification error can be obtained by solving:

$$\max_{\pi} \mathcal{L}(\mathcal{F}(A_{\pi(n)}(\cdots A_{\pi(1)}(x; \delta_{\pi(1)}) \cdots ; \delta_{\pi(n)})), y). \tag{2}$$

The Surrogate Image for Scheduling Optimization. Since $x_{\text{c-adv}}$ contains merely one attack perturbation at each iteration, using it alone is challenging to optimize the likelihood of other attacks in the relaxed scheduling matrix. To manage this issue, we adopt a surrogate composite adversarial image x_{surr} to relax the restriction and compute the loss for updating the scheduling matrix Z, i.e. by weighting each type of attack perturbation with its corresponding probability at each iteration. Therefore, we could optimize the scheduling matrix Z via maximizing the corresponding loss $\mathcal{L}(\mathcal{F}(x_{\text{surr}}), y)$. Given the attack pool Ω of n attacks, the surrogate image would be computed for n iterations. For each iteration i, the surrogate image is defined as:

$$x^i_{\text{surr}} = \sum_{j=1}^{n} z_{ij} \cdot A_j(x^{i-1}_{\text{surr}}; \delta_j)), \forall i \in \{1, \dots, n\}, \tag{3}$$

and $x^0_{\text{surr}} = x$. Let $\mathbf{A}^\top = \big(A_1, \dots, A_n\big)$ denotes a vector of all attack types in Ω. Consequently, after n iterations, the resulting surrogate image x^n_{surr} can be formulated into the following compositional form:

$$\begin{aligned} x^n_{\text{surr}} &= \mathbf{z}_n^\top \mathbf{A}(\cdots(\mathbf{z}_2^\top \mathbf{A}(\mathbf{z}_1^\top \mathbf{A}(x)))) \\ &= \mathbf{z}_n^\top \mathbf{A}(\cdots(\mathbf{z}_2^\top \mathbf{A}(\sum_{j=1}^{n} z_{1j} \cdot A_j(x; \delta_j)))) \\ &= \mathbf{z}_n^\top \mathbf{A}(\cdots(\mathbf{z}_2^\top \mathbf{A}(x^1_{\text{surr}}))) \\ &= \mathbf{z}_n^\top \mathbf{A}(\cdots(x^2_{\text{surr}})). \end{aligned} \tag{4}$$

How to Learn Optimal Attack Order? Learning an optimal attack order expressed by the scheduling matrix $Z^\star$ is originally a combinatorial optimization problem to solve the best column and row permutation of a scheduling matrix. Sinkhorn and Knopp proved that any positive square matrix could be turned into a doubly stochastic matrix by alternately normalizing the rows and columns of it [34]. Furthermore, Mena et al. theoretically showed how to extend the Sinkhorn normalization to learn and determine the optimal permutation matrix [25]. Similarly, in our problem, optimizing the attack order over a doubly stochastic matrix Z can be cast as a maximization problem, where the feasible solution set is convex. With the surrogate composite adversarial example x_{surr}, the updating process of the scheduling matrix Z^t for iteration t can be formulated as:

$$Z^t = \mathcal{S}\big(\exp(Z^{t-1} + \frac{\partial \mathcal{L}(\mathcal{F}(x_{\text{surr}}), y)}{\partial Z^{t-1}})\big), \tag{5}$$

where $\mathcal{S}$ (Sinkhorn normalization) can be done in a limited number of iterations [35]. Here, we fixed the iteration as 20 steps. After deriving an updated scheduling matrix, we utilize the Hungarian assignment algorithm to obtain the updated order assignment function $\pi_t(\cdot)$, as shown in Eq. 6:

$$\pi_t(j) := \arg\max \mathbf{z}_j, \forall j \in \{1, \dots, n\}. \tag{6}$$

3.2. The Component-wise PGD (Comp-PGD)

Upon addressing the attack scheduling issue, we now move forward to elucidate the design of adversarial perturbation in each attack type (component) of our composite adversarial attacks. For most of the semantic perturbations, their parameters are of continuous value. Therefore, we propose to search the parameters of semantic attacks by gradient descent algorithm within each continuous semantic space. In particular, we showed how to optimize the parameters in the following five different semantic perturbations, including (i) hue, (ii) saturation, (iii) brightness, (iv) contrast, and (v) rotation. We extend the iterative gradient sign method [17] to optimize our semantic perturbations for T iterations, which is defined as:

$$\delta_k^{t+1} = \text{clip}_{\epsilon_k}\big(\delta_k^t + \alpha \cdot \text{sign}(\nabla_{\delta_k^t} \mathcal{L}(\mathcal{F}(A_k(x;\delta_k^t)), y))\big), \tag{7}$$

where t denotes the iteration index, α is the step size of each iteration, $\nabla_{\delta_k^t}\mathcal{L}(\cdot)$ is the gradient of a loss function $\mathcal{L}$ with respect to the perturbation variable δ_k^t. Let $\epsilon_k = [\alpha_k, \beta_k]$, we denote the element-wise clipping operation $\text{clip}_{\epsilon_k}(z)$ as:

$$\text{clip}_{\epsilon_k}(z) = \text{clip}_{[\alpha_k,\beta_k]}(z) = \begin{cases} \alpha_k & \text{if } z < \alpha_k, \\ z & \text{if } \alpha_k \le z \le \beta_k, \\ \beta_k & \text{if } \beta_k < z. \end{cases}$$

Next, we elucidate each semantic attack. The concrete examples of each of them are shown in Appendix Hand the loss trace analysis of Comp-PGD is shown in Appendix C.

Hue. The hue value is defined on a color wheel in HSV color space, ranging from 0 to 2π. In hue attack (A_H), we define the perturbation interval of hue as $\epsilon_H = [\alpha_H, \beta_H]$, $-\pi \le \alpha_H \le \beta_H \le \pi$. Let $x_H = \text{Hue}(x)$ denote the hue value of an image x, the variation of hue value at step t is δ_H^t, and the initial variance δ_H^0 is chosen from ϵ_H uniformly. Then δ_H^t can be updated iteratively via Eq. 7, and the hue value of the perturbed image $x_{\text{c-adv}}^t = A_H(X;\delta_H^t)$ is:

$$x_H^t = \text{Hue}(x_{\text{c-adv}}^t) = \text{clip}_{[0,2\pi]}(x_H + \delta_H^t).$$

Saturation. Similar to hue value, saturation value determines the colorfulness of an image ranging from 0 to 1. Let $x_S = \text{Sat}(x)$ denote the saturation value of an image x. If $x_S \to 0$, the image becomes more colorless, resulting in a gray-scale image if $x_S = 0$. The perturbation interval of saturation is defined as $\epsilon_S = [\alpha_S, \beta_S]$, $0 \le \alpha_S \le \beta_S < \infty$. Let the perturbation factor of saturation value at step t is δ_S^t, and the initial factor δ_S^0 is chosen from ϵ_S uniformly. The saturation attack is to update the perturbation factor δ_S via Eq. 7, and the saturation value of the perturbed image $x_{\text{c-adv}}^t = A_S(X;\delta_S^t)$ is:

$$x_S^t = \text{Sat}(x_{\text{c-adv}}^t) = \text{clip}_{[0,1]}(x_S \cdot \delta_S^t).$$

Brightness and Contrast. Unlike hue and saturation, these values are defined on RGB color space (pixel space), and they determine the lightness, darkness, and brightness differences of images. In our implementation, we convert the images from $[0, 255]$ scale to $[0, 1]$. The perturbation interval of brightness and contrast is defined as $\epsilon_B = [\alpha_B, \beta_B]$, $-1 \le \alpha_B \le \beta_B \le 1$ and $\epsilon_C = [\alpha_C, \beta_C]$, $0 \le \alpha_C \le \beta_C < \infty$, respectively; the same, the initial perturbation δ_B^0 and δ_C^0 are chosen from ϵ_B and ϵ_C uniformly, and can update via Eq. 7. The perturbed image $x_{\text{c-adv}}^t$ under the brightness attack (A_B) and contrast attack (A_C) can be formulated as:

$$x_{\text{c-adv}}^t = \text{clip}_{[0,1]}(x + \delta_B^t) \;\text{ and }\; x_{\text{c-adv}}^t = \text{clip}_{[0,1]}(x \cdot \delta_C^t).$$

Rotation. This transformation aims to find a rotation angle such that the rotated image has a maximum loss. The rotation implementation is constructed by [30]. Given a square image x, let (i, j) denotes pixel position and (c, c) denotes the center position of x. Then the position (i', j') rotated by θ degree from (i, j) can be formulated as:

$$\begin{bmatrix} i' \\ j' \end{bmatrix} = \begin{bmatrix} \cos\theta \cdot i + \sin\theta \cdot j + (1-\cos\theta)\cdot c - \sin\theta \cdot c \\ -\sin\theta \cdot i + \cos\theta \cdot j + \sin\theta \cdot c + (1-\cos\theta)\cdot c \end{bmatrix}.$$

Here, we define the perturbation interval of rotation degree $\epsilon_R = [\alpha_R{}^\circ, \beta_R{}^\circ]$, $\alpha_R \le \beta_R$, $\alpha_R, \beta_R \in \mathbb{R}$. The perturbation degree at step t is δ_R^t, and the initial degree δ_R^0 is chosen from ϵ_R uniformly. Similarly, like the previous attack, the perturbation δ_R will be updated via Eq. 7.

3.3. Generalized Adversarial Training (GAT)

To harden the classifier against composite perturbations, we generalize the standard adversarial training approach with our proposed composite adversarial attack from Section 3.1. Our goal is to train a robust model $\mathcal{F}(\cdot)$ over a data distribution $(x, y) \sim \mathcal{D}$, and make it robust against composite perturbations in the perturbation boundary E. Existing adversarial training objectives such as the min-max loss [21] or TRADES loss [48] can be utilized in GAT. Here we use min-max training loss (Madry's loss) for illustration. The inner maximization in Eq. 8 is to generate $x_{\text{c-adv}}$ optimized using CAA within boundary E, and the outer minimization is for optimizing the model parameters $\theta_{\mathcal{F}}$.

$$\min_{\theta_{\mathcal{F}}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\max_{x_{\text{c-adv}} \in \mathcal{B}(x;\Omega;E)} \mathcal{L}(\mathcal{F}(x_{\text{c-adv}}), y) \right]. \tag{8}$$

For completeness, in Appendix Bwe summarize the flow of our proposed composite adversarial attacks with order scheduling and attack component optimization. In addition, the ablation study showing order-scheduling and Comp-PGD are essential can be found in Appendix E.

4. Experiments

In this section, we first elucidate the experimental settings and then present the performance evaluation and analysis against multiple composite attacks on two datasets: CIFAR-10 [15] and ImageNet [31]. Additional experimental results and implementation details are shown in Appendix G.

4.1. Experiment Setups

Datasets. We evaluated GAT on two different datasets: CIFAR-10 [15] and ImageNet [31]. CIFAR-10 consists of 60000 32*32 images, with 6000 images per class. There are 50000 training images and 10000 test images. ImageNet is a benchmark in image classification and object detection with 10 million images, including 1000 classes.

Attack Composition. There are many feasible combinations of threats can be utilized in the evaluation; we discuss two attack combinations here, *semantic attacks* and *full attacks*, with two scheduling strategies. Semantic attacks consist of a combination of five semantic perturbations, including *Hue*, *Saturation*, *Rotation*, *Brightness* and *Contrast* attacks. For full attacks, one can generate examples with *all five semantic attacks* and ℓ_∞ *attack*. We consider different order scheduling strategies: *scheduled* and *random*. That is, we can either schedule the order by the aforementioned scheduling algorithm in Sec. 3.1, or randomly shuffle an attack order when launching attacks for generating the corresponding composite adversarial examples. Furthermore, we also present the results of a variety of attack compositions for analysis (see Appendix F)and discuss the difference between separately/jointly optimizing the attack parameters in Appendix D.

Comparative Training Methods. We compare our GAT with several baseline adversarial training models on both datasets using two different model backbones: ResNet50 [10] and WideResNet34 [47]. The comparative methods are summarized in **Baseline Model Details** below. For GAT, we train our models via finetuning on the ℓ_∞-robust pretrained model for both CIFAR-10 and ImageNet and use the min-max loss in Eq. 8 [21]. Two ordering modes were adopted in GAT: random order (*GAT-f*) and scheduled order (*GAT-fs*). We also found that training from scratch using GAT is unstable due to the consideration of multiple perturbation threats (see Appendix A)

Baseline Model Details. In summary below, we use symbols to mark the model backbones. Here, † denotes models in ResNet50 [10] architecture and $*$ denotes models in WideResNet34 [47]. The baseline models are obtained from top-ranked models of the same architecture in RobustBench [5].

- **Normal**†**/Normal***: Standard training.
- **Madry**$^\dagger_\infty$: ℓ_∞ adversarial training in [21].
- **Trades**$^*_\infty$: ℓ_∞ adversarial training in [48].
- **FAT**$^*_\infty$: [49] uses friendly adversarial data that are confidently misclassified for adversarial training.
- **AWP**$^*_\infty$: [44] injects the worst-case weight perturbation during adversarial training to flatten the weight loss landscape.
- **PAT**$^\dagger_{self}$, **PAT**$^\dagger_{alex}$: Two adversarial training models based on the perceptual distance (LPIPS), two models differ: ResNet50 (*self*) and AlexNet (*alex*) [20].
- **Fast-AT**†: Computationally efficient ℓ_∞ adversarial training in [42].

Training & Evaluation Settings. We adopt the whole training set on both CIFAR-10 and ImageNet for model training. In every training iterative step, the images in each batch share the same attack order. Besides, the Comp-PGD is applied on each image, where we set the iteration-update step T as ten steps of each attack component for evaluation and seven steps for GAT. During the training of GAT, we apply every attack component on the input without the *early-stopped* option to ensure the model could learn all attack components which have been launched. Furthermore, we evaluate two different order scheduling settings: *random/scheduled* during GAT on CIFAR-10. Since both ordering mechanisms provide competitive robust models, therefore, we only use random ranking when training GAT on ImageNet, considering the training efficiency. As mentioned in Sec. 4.1, GAT utilizes a pre-trained model for fine-tuning to make the composite adversarial training more efficient than training from scratch. Different from the training phase of GAT, during the evaluation, we allow CAA to trigger the *early-stop* option when the attack is successful, which can help us improve the attack success rate and reduce the computational cost. Further discussion and comparison between different training settings of GAT, including using TRADES/Madry loss and fine-tuning/training from scratch, are given in Appendix A.

Computing Resources and Code. For CIFAR-10, we train models on ResNet50 and WideResNet34 with SGD for 150 epochs. The training of GAT-f takes about 16 hours (ResNet50) and 28 hours (WideResNet34), and GAT-fs takes about 28 hours (ResNet50) and 55 hours (WideResNet34) on 8 Nvidia Tesla V100 GPUs. For ImageNet, we train ResNet50 with SGD for 100 epochs and about three days on 64 Nvidia Tesla V100 GPUs. The implementation is built with PyTorch [28].

Evaluation Metrics. We report the models' Clean and Robust Accuracy (RA, %) against multiple composite adversarial attacks. The RA aims to evaluate the model accuracy toward the fraction of perturbed examples retrieved from the test set which is correctly classified. We also provide the attack success rate (ASR, %) in Appendix G, in which the higher indicates the stronger attack.

Training	Clean	Three attacks			Semantic attacks		Full attacks	
		CAA_{3a}	CAA_{3b}	CAA_{3c}	Rand.	Sched.	Rand.	Sched.
Normal†	95.2	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	59.7 ± 0.2	44.2 ± 0.5	0.0 ± 0.0	0.0 ± 0.0
Madry$^\dagger_\infty$	87.0	30.8 ± 0.2	18.8 ± 0.5	19.1 ± 0.3	31.5 ± 0.2	21.3 ± 0.2	10.8 ± 0.2	3.7 ± 0.2
PAT$^\dagger_{self}$	82.4	20.9 ± 0.1	11.9 ± 0.5	17.9 ± 0.3	28.9 ± 0.3	17.5 ± 0.3	9.1 ± 0.3	2.5 ± 0.3
PAT$^\dagger_{alex}$	71.6	20.7 ± 0.3	12.5 ± 0.2	16.5 ± 0.4	23.4 ± 0.3	12.2 ± 0.4	10.3 ± 0.1	2.5 ± 0.2
GAT-f†	**82.3**	**39.9 ± 0.1**	**33.3 ± 0.1**	**28.9 ± 0.2**	**69.9 ± 0.1**	**66.0 ± 0.1**	**30.0 ± 0.4**	**18.8 ± 0.3**
GAT-fs†	**82.1**	**43.5 ± 0.1**	**36.6 ± 0.1**	**32.5 ± 0.1**	**69.9 ± 0.2**	**66.6 ± 0.1**	**32.3 ± 0.8**	**21.8 ± 0.3**
Normal*	94.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	46.0 ± 0.4	29.9 ± 0.5	0.0 ± 0.0	0.0 ± 0.0
Trades$^*_\infty$	84.9	30.0 ± 0.3	19.8 ± 0.6	10.1 ± 0.5	16.6 ± 0.2	8.1 ± 0.5	5.8 ± 0.3	1.5 ± 0.2
FAT$^*_\infty$	88.1	29.8 ± 0.4	17.1 ± 0.4	12.8 ± 0.6	18.7 ± 0.2	9.8 ± 0.5	6.1 ± 0.1	1.5 ± 0.1
AWP$^*_\infty$	85.4	34.2 ± 0.2	23.2 ± 0.2	11.1 ± 0.4	15.6 ± 0.2	7.9 ± 0.2	5.9 ± 0.0	1.7 ± 0.2
GAT-f*	**83.4**	**40.2 ± 0.1**	**34.0 ± 0.1**	**30.7 ± 0.4**	**71.6 ± 0.1**	**67.8 ± 0.2**	**31.2 ± 0.4**	**20.1 ± 0.3**
GAT-fs*	**83.2**	**43.5 ± 0.1**	**36.3 ± 0.1**	**32.9 ± 0.4**	**70.5 ± 0.1**	**66.7 ± 0.3**	**32.2 ± 0.7**	**21.9 ± 0.7**

Table 1. Comparison of accuracy (%) on CIFAR-10. We combine different types of three attacks (CAA_3) with scheduled ordering: CAA_{3a}: (Hue, Saturation, ℓ_∞), CAA_{3b}: (Hue, Rotation, ℓ_∞), CAA_{3c}: (Brightness, Contrast, ℓ_∞), on CIFAR-10

Training	Clean	Three attacks			Semantic attacks		Full attacks	
		CAA_{3a}	CAA_{3b}	CAA_{3c}	Rand.	Sched.	Rand.	Sched.
Normal†	76.1	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	31.2 ± 0.4	20.6 ± 1.0	0.0 ± 0.0	0.0 ± 0.0
Madry$^\dagger_\infty$	62.4	13.9 ± 0.4	9.2 ± 0.2	16.2 ± 0.8	14.0 ± 0.1	9.0 ± 0.0	7.1 ± 0.1	2.8 ± 0.2
Fast-AT$^\dagger_\infty$	53.8	9.5 ± 0.3	5.5 ± 0.1	11.4 ± 0.8	6.3 ± 0.1	3.6 ± 0.1	3.1 ± 0.1	1.0 ± 0.1
GAT-f†	60.0	**19.2 ± 1.0**	**18.9 ± 1.4**	**18.4 ± 0.4**	**43.5 ± 1.9**	**38.9 ± 2.0**	**18.5 ± 0.5**	**11.8 ± 0.1**

Table 2. Comparison of accuracy (%) on ImageNet. (CAA$_{3a,3b,3c}$: same combination as Table 1)

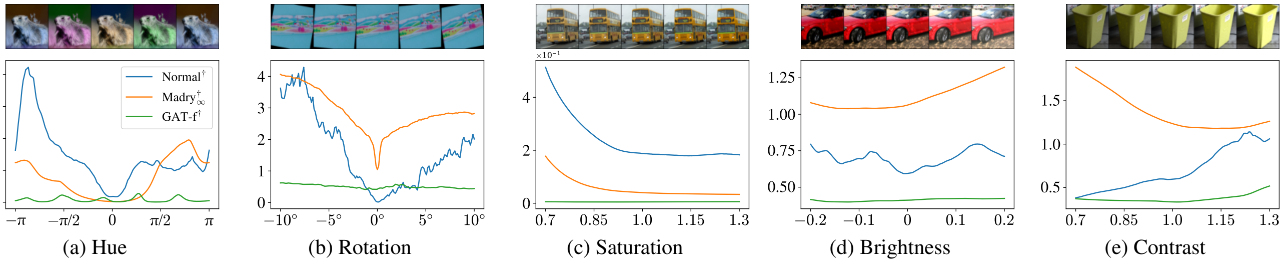
4.2. Performance Evaluation

The experimental results are shown in Table 1 (CIFAR-10) and Table 2 (ImageNet). On CIFAR-10, *GAT-fs* and *GAT-f* show competitive results. Both of them outperform all other baselines by a significant margin. For semantic attacks, the RA increases by 45% $\sim$ 60% on CIFAR-10, and 28% $\sim$ 37% on ImageNet. For full attacks, the RA increases by 15% $\sim$ 27% on CIFAR-10, and 9% $\sim$ 15% on ImageNet. Nonetheless, the RA against three multiple threats with three different combinations, our proposed GAT keeps outperforming other baselines and shows the highest robustness of others. The comparison between GAT-f and GAT-fs demonstrates that GAT-fs can obtain higher RA against full attacks. However, the result also suggests a trade-off between the robustness of ℓ_∞ and semantic attacks.

Besides adversarial training models, we empirically observe that the RA of models with standard training has a degraded performance of 20% $\sim$ 31% on ImageNet data under semantic attacks (without ℓ_∞ attack). However, while ℓ_∞ attack is involved in the full attacks or other multiple threats (e.g., three attacks in Tables 1 and 2), the models with only standard training are unable to resist these kinds of composite semantic perturbations, and the RA drops dramatically to 0%.

4.3. Analysis, Discussion, and Visualization

Robust Accuracy vs. Number of Attacks and Their combinations. We conduct an ablation study to show that the number of attacks and their combinations can hugely affect robust accuracy, illustrating the importance of attack ordering and the new insights into robustness through our proposed composite adversarial examples. Fig. 1b already demonstrates that our model is the most resilient to composite adversarial examples consisting of different numbers of attacks, in terms of attaining the lowest attack success rate in the test set that each model initially correctly classified. Furthermore, Table 3 shows that as the number of attacks increases (CAA_1 to CAA_6), the RA of our proposed GAT consistently outperforms all other models. Specifically, GAT outperforms other baselines by up to 35%. Although the standard model (Normal†) has the advantage of higher clean-ing accuracy, it is still not resistant to semantic and various composite adversarial perturbations. Results of *three attacks* in Tables 1 and 2 demonstrate the effect of different combinations when the number of attacks is fixed. Comparing GAT with others on both CIFAR-10 and ImageNet, the result shows that *GAT-f* is more robust than all baselines under three different attacks by 9% $\sim$ 23%. On ImageNet, *GAT-f* also outperforms those baselines. For more experimental results, including single attacks, Auto-attack, two-component



(a) Hue (b) Rotation (c) Saturation (d) Brightness (e) Contrast

Figure 3. Loss landscape of selected examples when performing five different semantic attacks under models produced by different adversarial training approaches. The blue and orange color curves represent standard and ℓ_∞ robust model respectively; the green color curve represents GAT-f model.

Training	CAA_1	CAA_2	CAA_3	CAA_4	CAA_5	CAA_6
Normal†	50.9	45.8	33.4	25.8	21.1	0.0
Madry$^\dagger_\infty$	38.1	33.9	21.9	14.4	9.0	2.8
Fast-AT$^\dagger_\infty$	27.8	23.9	12.7	7.0	3.6	1.0
GAT-f†	**51.0**	**48.2**	**44.5**	**42.2**	**38.9**	**11.8**

Table 3. Comparison of RA (%) on four adversarial training approaches against six different *CAAs* on ImageNet. CAA_1: (Hue), CAA_2: (Hue, Saturation), CAA_3: (Hue, Saturation, Rotation), CAA_4: (Hue, Saturation, Rotation, Brightness), CAA_5: (Hue, Saturation, Rotation, Brightness, Contrast), CAA_6: (Hue, Saturation, Rotation, Brightness, Contrast, ℓ_∞)

attacks, and other results on other datasets (e.g., SVHN), please refer to Appendix G.

Effectiveness of Random vs. Scheduled Ordering. We evaluated the effectiveness of random versus scheduled ordering by conducting pairwise t-tests. Ten experiments were conducted with different initializations, and the experimental results on CIFAR10/Full-attack demonstrated that the robust accuracy of the *scheduled* ordering was significantly lower than that of the *random* ordering (p-value $<$ .001 for all models).

Inadequacies of Current Adversarial Robustness Assessments. Existing methods for evaluating adversarial robustness, which only considers perturbations in ℓ_p-ball, may be incomplete and biased. To investigate this issue, we compared the rankings of the top ten models on the RobustBench dataset (CIFAR-10, ℓ_∞) [5]. We found that the rankings between Auto-Attack and CAA had a low correlation, suggesting the need for more comprehensive assessments. Specifically, we computed the Spearman's rank correlation coefficient between Auto-Attack and CAA (rand. & sched.) for semantic and full attacks, which yielded values of 0.16 (rand. *vs.* sched.) and 0.36 (rand. *vs.* Auto) and 0.38 (sched. *vs.* Auto), respectively. These findings underscore the importance of developing new methods that can more accurately evaluate adversarial robustness.

Visualization of Loss Landscape. To gain a deeper understanding of why our proposed approach leads to significant improvements in adversarial robustness, we visualized the loss landscape of a single semantic attack under three different models: standardly trained ResNet50 (Normal†), ResNet50 with ℓ_∞-robust training (Madry$^\dagger_\infty$), and our proposed GAT approach (*GAT-f*†), see Fig. 3. We plotted the cross-entropy loss of selected samples for each model, sweeping over the semantic perturbation space within a designated interval. We empirically observe that across five different single semantic attacks, the curves (green) generated by GAT were much smoother, flatter, and lower than those produced by the other models. We believe that this phenomenon sheds light on the effectiveness of our proposed approach, which can indeed train a model robust to the composite adversarial perturbations.

5. Conclusion

In this paper, we proposed GAT, a generic approach for enhancing the robustness of deep learning models to composite semantic perturbations, with the ultimate goal of preparing classifiers for the real world. Our approach is based on a unique design of attack order scheduling for multiple perturbation types and the optimization of each attack component. This further enables GAT to achieve state-of-the-art robustness against a wide range of adversarial attacks, including those in ℓ_p norms and semantic spaces. Evaluated on CIFAR-10 and ImageNet datasets, our results demonstrate that GAT achieves the highest robust accuracy on most composite attacks by a large margin, providing new insights into achieving compositional adversarial robustness. We believe our work sheds new light on the frontiers of realistic adversarial attacks and defenses.

6. Acknowledgement

Part of this work was done during L. Hsiung's visit to IBM Thomas J. Watson Research Center. The work of L. Hsiung and T.-Y. Ho was supported by the Ministry of Science and Technology, Taiwan (MOST 111-2218-E-005-006-MBK). We thank National Center for High-performance Computing (NCHC) in Taiwan for providing computational and storage resources.

References

[1] Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and D. A. Forsyth. Unrestricted adversarial examples via semantic manipulation. In *International Conference on Learning Representations*, 2020. 3

[2] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018. 2

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57, 2017. 2

[4] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. EAD: elastic-net attacks to deep neural networks via adversarial examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10–17, 2018. 2

[5] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 6, 8

[6] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 2

[7] Isaac Dunn, Laura Hanu, Hadrien Pouget, Daniel Kroening, and Tom Melham. Evaluating robustness to context-sensitive feature perturbations of different granularities. *arXiv preprint arXiv:2001.11055*, 2020. 3

[8] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning*, pages 1802–1811. PMLR, 2019. 3

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 6

[11] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1614–1619, 2018. 3

[12] Matt Jordan, Naren Manoj, Surbhi Goel, and Alexandros G Dimakis. Quantifying perceptual distortion of adversarial examples. *arXiv preprint arXiv:1902.08265*, 2019. 3

[13] Ameya Joshi, Amitangshu Mukherjee, Soumik Sarkar, and Chinmay Hegde. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4773–4783, 2019. 3

[14] Daniel Kang, Yi Sun, Dan Hendrycks, Tom Brown, and Jacob Steinhardt. Testing robustness against unforeseen adversaries. *arXiv preprint arXiv:1908.08016*, 2019. 3

[15] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. 6

[16] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 4

[17] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *ICLR Workshop*, 2017. 5

[18] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *International Conference on Learning Representations*, 2017. 3

[19] Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 3

[20] Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. In *ICLR*, 2021. 2, 3, 6

[21] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018. 1, 2, 3, 5, 6

[22] Pratyush Maini, Eric Wong, and Zico Kolter. Adversarial robustness against the union of multiple perturbation models. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6640–6650. PMLR, 13–18 Jul 2020. 3

[23] Xiaofeng Mao, Yuefeng Chen, Shuhui Wang, Hang Su, Yuan He, and Hui Xue. Composite adversarial attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8884–8892, May 2021. 3

[24] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12032–12041, 2022. 3

[25] Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. In *International Conference on Learning Representations*, 2018. 4

[26] Jeet Mohapatra, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Towards verifying robustness of neural networks against a family of semantic perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[27] James R. Munkres. Algorithms for the assignment and transportation problems. *Journal of The Society for Industrial and Applied Mathematics*, 10:196–210, 1957. 4

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information*

Processing Systems 32, pages 8024–8035. Curran Associates, Inc., 2019. 6

[29] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *European Conference on Computer Vision*, pages 19–37. Springer, 2020. 3

[30] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3674–3683, 2020. 5

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 6

[32] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and Andrea Cavallaro. Colorfool: Semantic adversarial colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1151–1160, 2020. 3

[33] Yash Sharma and Pin-Yu Chen. Attacking the Madry defense model with L_1-based adversarial examples. *ICLR Workshop*, 2018. 3

[34] Richard Sinkhorn. A relationship between arbitrary positive matrices and stochastic matrices. *Canadian Journal of Mathematics*, 18:303–306, 1966. 4

[35] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967. 4

[36] David Stutz, Matthias Hein, and Bernt Schiele. Confidence-calibrated adversarial training: Generalizing to unseen attacks. In *International Conference on Machine Learning*, pages 9155–9166. PMLR, 2020. 3

[37] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *International Conference on Learning Representations*, 2014. 1

[38] Florian Tramer and Dan Boneh. Adversarial training and robustness for multiple perturbations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 3

[39] Jingkang Wang, Tianyun Zhang, Sijia Liu, Pin-Yu Chen, Jiacen Xu, Makan Fardad, and Bo Li. Adversarial attack generation empowered by min-max optimization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16020–16033. Curran Associates, Inc., 2021. 3

[40] Shuo Wang, Shangyu Chen, Tianle Chen, Surya Nepal, Carsten Rudolph, and Marthie Grobler. Generating semantic adversarial examples via feature manipulation. *arXiv preprint arXiv:2001.02297*, 2020. 3

[41] Yajie Wang, Shangbo Wu, Wenyi Jiang, Shengang Hao, Yu-an Tan, and Quanxin Zhang. Demiguise attack: Crafting invisible semantic adversarial perturbations with perceptual similarity. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3125–3133. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track. 3

[42] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020. 2, 6

[43] Eric Wong, Frank Schmidt, and Zico Kolter. Wasserstein adversarial examples via projected sinkhorn iterations. In *International Conference on Machine Learning*, pages 6808–6817. PMLR, 2019. 3

[44] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *NeurIPS*, 2020. 2, 6

[45] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations*, 2018. 3

[46] Zheng Yuan, Jie Zhang, and Shiguang Shan. Adaptive image transformations for transfer-based adversarial attack. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, page 1–17, Berlin, Heidelberg, 2022. Springer-Verlag. 3

[47] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016. 6

[48] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019. 1, 2, 3, 5, 6

[49] 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 *ICML*, 2020. 2, 6

[50] Dawei Zhou, Tongliang Liu, Bo Han, Nannan Wang, Chunlei Peng, and Xinbo Gao. Towards defending against adversarial examples via attack-invariant features. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12835–12845. PMLR, 18–24 Jul 2021. 3