Supplementary Material - 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/

A. Implementation Details

Training Phase. In the implementation of generalized adversarial training (GAT), we consider two model architectures, ResNet-50 [2] and WideResNet-34 [6], on CIFAR-10 dataset [3]; and ResNet-50 on ImageNet dataset [5]. For CIFAR-10, we set the maximum training epoch to 150 with the batch size 2048 and selected the model with the best evaluation test accuracy. The learning rate is set to 0.1 at the beginning and exponentially decays. We utilize the warming-up learning rate technique for the first ten epochs, which means the learning rate would linearly increase from zero to the preset value (0.1) in the first ten epochs. For ImageNet, we set the maximum training epoch to 100 with the batch size 1536 and selected the model with the best evaluation test accuracy. The learning rate is set to 0.1 at the beginning and exponentially decays by 0.1 every 30 epochs. Similarly, we utilize the warming-up learning rate technique for the first five epochs. We launched all threat models (full attacks) while training; for each batch, we utilized scheduled ordering for *GAT-fs* and random ordering for *GAT-f*. The iteration step T of each attack for Comp-PGD is set to 7, and the step size of attack A_k is set as $2.5 \cdot (\beta_k - \alpha_k)/2T$, where β_k and α_k are the values of perturbation intervals defined in Table A1.

	CIFAR-10, SVHN	ImageNet
Hue, ϵ_H	$-\pi \sim \pi$	
Saturation, ϵ_S	$0.7 \sim 1.3$	
Rotation, ϵ_R	$-10° \sim 10°$	
Brightness, ϵ_B	$-0.2 \sim 0.2$	
Contrast, ϵ_C	$0.7 \sim 1.3$	
ℓ_∞, ϵ_L	8/255	4/255

Table A1. Perturbation interval of each attack component

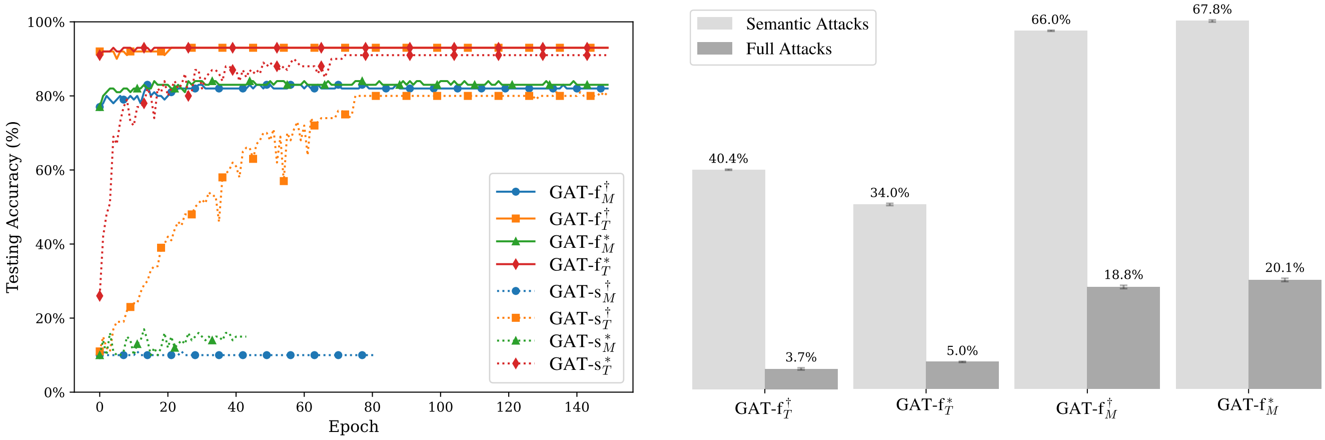
Testing Phase. To compare our GAT approach with other adversarial training baselines, we launch composite adversarial attacks (CAAs) of different numbers of attack types, including single attacks, two attacks, three attacks, all semantic attacks, and full attacks on each robust model. Furthermore, the iteration step T of each attack for Comp-PGD is set as 10, and the step size is the same as the training settings. In addition, the maximum iteration of *order scheduling* is designated as five, and we will launch the early-stop option in every update step while the CAA succeeds in attacking. Note that the ASR would slightly decrease ($\approx 2\%$) if the early-stop feature is disabled. This is likely due to the highly complex and non-convex loss landscape (Fig. A2); while the early-stop feature helps CAA maintain its attack efficiency.

Training Strategy. Our training process considers two training strategies: 1) training from scratch and 2) fine-tuning on ℓ_∞-robust models; two learning objectives: 1) Madry's loss [4] and 2) TRADES' loss [7]. Note that $x_{\text{c-adv}} \in \mathcal{B}(x;\Omega;E)$ denotes the composite adversarial example $x_{\text{c-adv}}$ is perturbed by attacks from Ω within the perturbation intervals E. The main difference between these two is shown in Eq. 1 and Eq. 2. That is, Eq. 2 encourages the natural error to be optimized in the first term; meanwhile, the robust error in the second regularization term could help minimize the distance between the prediction of natural samples and adversarial samples. Zhang et al. theoretically proved that this design of loss function could help the outputs of the model to be smooth [7].

$$\min_{\theta_{\mathcal{F}}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\max_{x_{\text{c-adv}} \in \mathcal{B}(x;\Omega;E)} \mathcal{L}_{ce}(\mathcal{F}(x_{\text{c-adv}}), y) \right] \tag{1}$$

$$\min_{\theta_{\mathcal{F}}} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}(\mathcal{F}(x), y) + \beta \cdot \max_{x_{\text{c-adv}} \in \mathcal{B}(x;\Omega;E)} \mathcal{L}(\mathcal{F}(x), \mathcal{F}(x_{\text{c-adv}})) \right] \tag{2}$$

As shown in Fig. A1a, we evaluate the clean test accuracy of GAT models in every epoch with different training settings, including using two architectures (ResNet-50 / WideResNet-34), two learning objectives, and two training strategies mentioned above. We empirically find the models using fine-tuning strategy (solid curves) can achieve higher clean test accuracy than most of the models training from scratch (dotted curves). Furthermore, we evaluate the robust test accuracy for these four models (see Fig. A1b). Under the semantic and full attacks, the models GAT-f$_M$ (fine-tuning with Madry's loss) achieve higher robust accuracy than GAT-f$_T$ (fine-tuning with TRADES loss). Hence, in the section of experimental results, we utilized the GAT models, which are trained with Madry's loss and fine-tuning on a ℓ_∞-robust model.



(a)

(b)

Figure A1. (a) The testing accuracy during generalized adversarial training on CIFAR-10. The models differ in different training scenarios; the lower script T denotes that the model using *TRADES*' loss [7] for training, and M for *Madry*'s loss [4]. The upper script † denotes the model using ResNet50 [2] backbone and $*$ is for WideResNet34 [6]. (b) The robust accuracy (%) of our GAT fine-tuned models under semantic and full attacks.

B. Algorithm of the Composite Adversarial Attack (CAA)

Algorithm 1: Composite Adversarial Attack

```
Input: classifier $\mathcal{F}(\cdot)$, input $x$, label $y$, attack space $\pi = \{A_1, \dots, A_n\}$, attack order schedule, scheduling iterations
       $M$, perturbation intervals $\{\epsilon_k\}_{k=1}^{n}$, Comp-PGD steps $T$
Output: Composite adversarial examples $x_{\text{c-adv}}$
# Initialization
$\delta_1^0, \dots, \delta_n^0 \leftarrow$ initial perturbation
if schedule == scheduled then
    $\mathcal{Z}^0, \pi_0 \leftarrow$ scheduling matrix and order assignment initialization
else
    $\pi_0 \leftarrow$ initial order assignment (random / fixed)
# Iteration of attack order scheduling
for $i \in \{1, \dots, M\}$ do
    # Applying attacks in order
    for $k \in \{1, \dots, n\}$ do
        $A_* = A_{\pi_i(k)}$ , $\delta_*^0 = \delta_{\pi_i(k)}^0$ , $\epsilon_* = \epsilon_{\pi_i(k)}$
        $x_{\text{c-adv}}^k \leftarrow A_*(x_{\text{c-adv}}^{k-1}; \delta_*^0)$
        # Iteration of Comp-PGD
        for $t \in \{1, \dots, T\}$ do
            if $\mathcal{F}(x_{\text{c-adv}}^k) \neq y$ then
                return $x_{\text{c-adv}}^k$    # Early stop option
            else
                $\delta_*^t = \text{clip}_{\epsilon_*}(\cdot; x_{\text{c-adv}}^k; \delta_*^{t-1})$ by Eq. 7
                $x_{\text{c-adv}}^k \leftarrow A_*(x_{\text{c-adv}}^{k-1}; \delta_*^t)$
    # Resetting the attack order
    if schedule == random then
        $\pi_{i+1} \leftarrow$ Shuffle a new order
    else if schedule == scheduled then
        # Optimize scheduling order $Z$
        $x_{\text{surr}} = \mathbf{z}_n^\top \mathbf{A}(\cdots(\mathbf{z}_2^\top \mathbf{A}(\mathbf{z}_1^\top \mathbf{A}(x))))$    # Compute the surrogate composite adversarial example by Eq. 4.
        $Z^t = \mathcal{S}\big(\exp(Z^{t-1} + \partial\mathcal{L}(\mathcal{F}(x_{\text{surr}}), y)/\partial Z^{t-1})\big)$    # Updating the scheduling matrix by Eq. 5.
        $\pi_{i+1}(j) := \arg\max \mathbf{z}_j, \forall j \in \{1, \dots, n\}$    # Update the attack order assignment by Eq. 6.
return $x_{\text{c-adv}}$
```

C. The Loss Trace Analysis of Component-wise PGD (Comp-PGD)

To demonstrate the effectiveness of Comp-PGD, in Fig. A2, we visualize the update process of Comp-PGD when performing a single semantic attack on the WideResnet-34 model. We uniformly sample 20 start points for each attack and update δ_k using Comp-PGD by these initial points. The red margins of each sub-figure in Fig. A2 represent the margin of successful attack by our samples. The endpoints of the loss trace show obviously that Comp-PGD indeed can help search for the worst case by maximizing the loss during each attack.

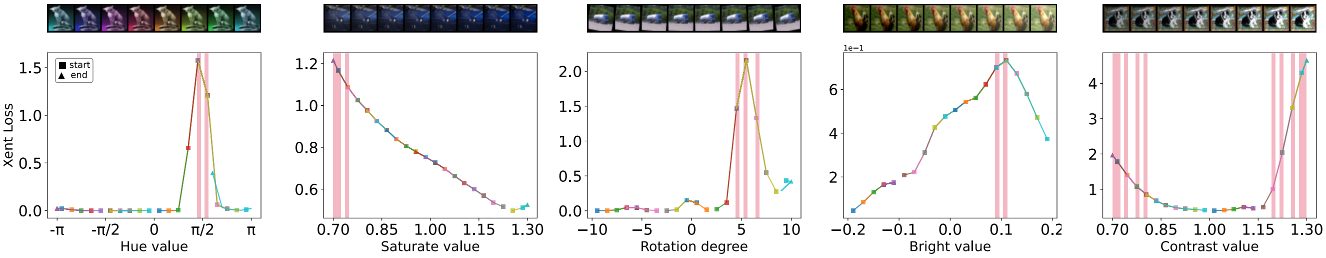


Figure A2. Component-wise PGD process of the single semantic attack.

D. Ablation Study: Attack Components' Optimization

D.1. Why Separately Optimize the Attack Parameters? (Comp-PGD vs. Ensemble-PGD)

In this paper, we used Comp-PGD to optimize the individual attack component. On the other hand, one can also optimize all attack components simultaneously given an attack order, for which we call *Ensemble-PGD*. Specifically, CAA can jointly optimize the attack parameters for an attack chain *at a chosen fixed attack order*. In this regard, we repeated the same experiments on CIFAR-10 but considered optimizing the attack parameters *simultaneously* instead of *sequentially*. The results show that Ensemble-PGD does not provide better attack capacity (see Table A2) than Comp-PGD (see Table A11). We provide the experimental results in Attack Success Rate (ASR), as it represents the strength of the attack (higher means a more vigorous attack). Although GAT approaches still outperform other baselines in defending against all threats, the results showed that Ensemble-PGD generally has *lower* attack performance than Comp-PGD. This is probably due to the fact that the number of the variables for optimizing in Ensemble-PGD is higher than that of Comp-PGD (in each sequential step), making the optimization process harder to achieve similar results.

		Three attacks			Semantic attacks		Full attacks	
Training	Clean	CAA_{3a}	CAA_{3b}	CAA_{3c}	Rand.	Sched.	Rand.	Sched.
Normal†	0.0	96.4	91.4	99.4	25.1	30.9	80.7	91.5
Madry$^\dagger_\infty$	0.0	43.2	54.7	57.9	37.6	46.3	57.7	72.0
PAT$^\dagger_{self}$	0.0	50.1	60.3	67.8	42.1	50.9	63.3	74.3
PAT$^\dagger_{alex}$	0.0	45.5	56.3	63.7	50.2	57.1	64.1	73.6
GAT-f†	0.0	**50.9**	**50.8**	**63.0**	**7.4**	**9.0**	**39.7**	**56.2**
GAT-fs†	0.0	**47.3**	**50.0**	**52.7**	**7.7**	**9.3**	**40.3**	**52.5**
Normal*	0.0	96.8	89.9	99.6	32.3	38.8	83.9	87.6
Trades$^*_\infty$	0.0	40.7	52.7	66.2	48.5	58.3	64.2	74.3
FAT$^*_\infty$	0.0	42.1	57.6	64.9	47.9	59.6	65.8	76.6
AWP$^*_\infty$	0.0	37.2	50.1	62.5	50.0	58.3	64.6	77.1
GAT-f*	0.0	**47.0**	**49.0**	**52.8**	**6.8**	**8.8**	**41.1**	**54.2**
GAT-fs*	0.0	**46.3**	**48.8**	**52.9**	**6.9**	**8.7**	**40.0**	**53.9**

Table A2. Attack success rate (%) of using Ensemble-PGD to perform CAA on CIFAR-10.

D.2. Why Not Optimize Attack Power by Grid Search? (Comp-PGD vs. Grid Search)

It is intuitive to optimize the attack parameters (levels) in a brute-force way, i.e., *grid search*. However, doing so would exponentially increase the computational cost as the number of attacks increases. We conduct an experiment to compare the attack success rate (ASR, %) between the Grid-Search attack and our proposed CAA. We include all types of semantic attacks (Hue, Saturation, Rotation, Brightness, and Contrast) in this experiment. Also, since there are $N!$ kinds of attack orders for N attacks, for simplicity, we chose only one attack order here and utilized the same attack order in CAA (fixed).

As shown in Fig. A3, the results demonstrated that CAA is obviously stronger than grid search, with a significantly lower computational cost. The results also indicate that CAA is more valuable than grid-search-based optimization, as CAA consistently achieves a higher ASR. This is because, in grid search, it could only look into the discrete attack value space; clearly, it would need to increase spatial density (grid numbers) to obtain a higher attack success rate. To be more specific, given the grid numbers K (uniformly sampled points in each attack space), the attack complexity of Grid-Search Attack is $\mathcal{O}(K^N)$; and the attack complexity of CAA (fixed order) is $\mathcal{O}(N \cdot T \cdot R)$, where T is the optimization steps for Comp-PGD, and R is the number of restarts. That is, we allow CAA to optimize each attack with R different starting points. In our experiment, since CAA could search for the optimal attack value by gradient-based search, we need only five restart points (R) and ten steps for Comp-PGD optimization (T) to outperform the grid-search-based strategy. In this scenario, the attack complexity of Grid-Search Attack is higher than CAA (since $\mathcal{O}(K^N) > \mathcal{O}(N \cdot K^2) > \mathcal{O}(N \cdot T \cdot R)$, given $T, R \leq K$).

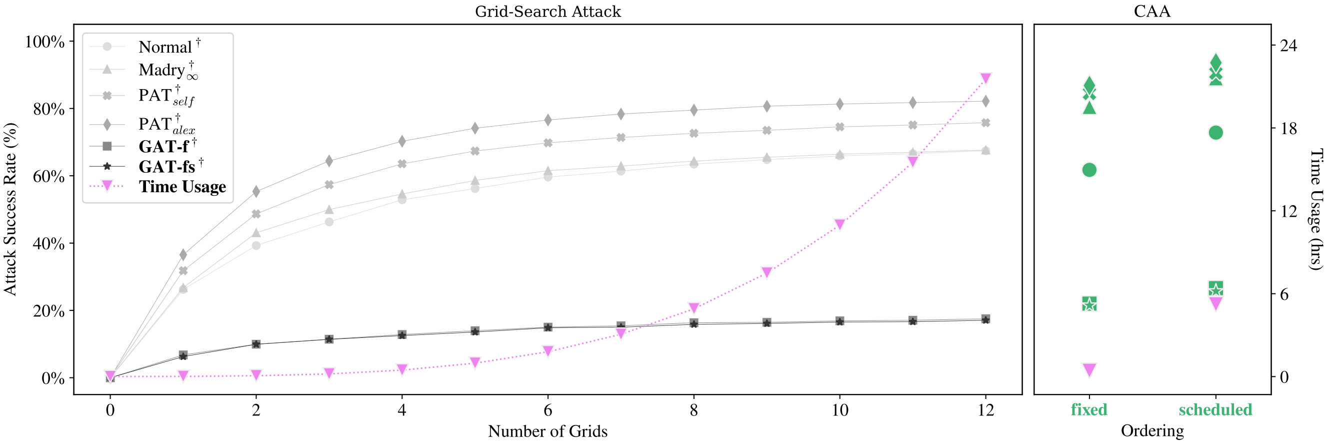


Figure A3. Comparison of attack success rate between Grid-Search Attack and CAA.

E. Ablation Study: Order Scheduling and Comp-PGD Are Essential to Strengthen GAT

To further verify that our scheduling mechanism and Comp-PGD play essential roles in CAA while doing GAT, we remove the order scheduling feature and Comp-PGD but pre-generate training data by adding random semantic perturbations on the CIFAR-10 training set, referring to RSP-10. That is, RSP-10 is generated in random attack ordering and random attack parameters on CIFAR-10. We then performed regular adversarial training on RSP-10 to obtain the robust models [7], including from-scratch and fine-tuning. Table A3 listed the robust accuracy of three such robust models under three attacks, semantic attacks and full attacks. The results show that GAT still outperforms other baselines for up to 27%/54%/25% in three/semantic/full attacks, demonstrating that order scheduling and Comp-PGD are essential to harden GAT to derive a robust model.

		Three attacks			Semantic attacks		Full attacks	
Training	Clean	CAA_{3a}	CAA_{3b}	CAA_{3c}	Rand.	Sched.	Rand.	Sched.
RSP*	84.9	17.9 ± 0.7	11.6 ± 0.9	5.9 ± 0.6	22.8 ± 0.2	12.6 ± 0.0	6.7 ± 0.3	1.6 ± 0.2
RSP-N*	88.1	18.8 ± 0.4	11.9 ± 0.3	8.8 ± 0.5	39.4 ± 0.5	28.5 ± 0.3	10.2 ± 0.2	3.9 ± 0.3
RSP-T*	85.4	38.3 ± 0.2	28.4 ± 0.2	21.2 ± 0.5	54.6 ± 0.3	47.5 ± 0.7	20.4 ± 0.7	9.8 ± 1.0
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**

Note. RSP*: AT from scratch; RSP-N*: AT, fine-tuned on Normal*; RSP-T*: AT, fine-tuned on Trades$^{*}_{\infty}$

Table A3. Robust accuracy (%) of models trained with simulated CAA samples.

F. Sensitive Analysis and Additional Discussions

F.1. The Attack Order Matters! The Two-attack Experiments

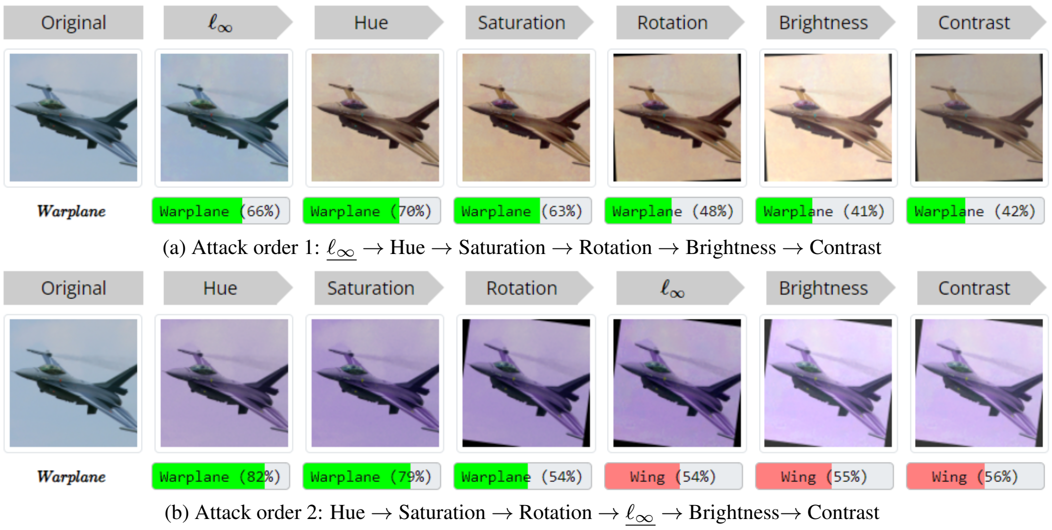
We conduct an analysis on different *order* types under two attacks to demonstrate the influence of order on CAA. As shown in Table A4, we list the attack success rate (ASR) of two attacks with different orders ($\ell_\infty \to$ *semantic* attack / *semantic* attack $\to \ell_\infty$) on GAT and other baseline models. The results show that most baselines are more fragile to the CAA with a semantic attack launched first than the attack with ℓ_∞ first. Furthermore, *GAT-f* has the smallest ASR change when alternating the order, indicating that GAT helps improve the robustness when the attack order is changed.

	2 attacks (Semantic $\to \ell_\infty$)				
Training	Hue $\to \ell_\infty$	Saturation $\to \ell_\infty$	Rotation $\to \ell_\infty$	Brightness $\to \ell_\infty$	Contrast $\to \ell_\infty$
Normal†	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
Madry$^\dagger_\infty$	72.3 ± 2.0	52.4 ± 0.7	60.6 ± 0.3	56.5 ± 0.5	58.2 ± 0.7
Fast-AT$^\dagger_\infty$	76.7 ± 1.8	56.7 ± 0.6	66.6 ± 0.1	62.4 ± 0.7	63.9 ± 0.7
GAT-f†	60.9 ± 2.4	59.6 ± 0.8	60.8 ± 0.6	59.7 ± 0.6	64.2 ± 0.5
	2 attacks ($\ell_\infty \to$ Semantic)				
Training	$\ell_\infty \to$ Hue	$\ell_\infty \to$ Saturation	$\ell_\infty \to$ Rotation	$\ell_\infty \to$ Brightness	$\ell_\infty \to$ Contrast
Normal†	100.0 ± 0.0	100.0 ± 0.0	99.9 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
Madry$^\dagger_\infty$	64.6 ± 0.5 (**7.7**$\downarrow$)	49.1 ± 0.5 (**3.3**$\downarrow$)	52.5 ± 0.1 (**8.0**$\downarrow$)	50.9 ± 0.4 (**5.7**$\downarrow$)	48.8 ± 0.5 (**9.4**$\downarrow$)
Fast-AT$^\dagger_\infty$	71.5 ± 0.6 (**5.2**$\downarrow$)	54.1 ± 0.6 (**2.6**$\downarrow$)	60.0 ± 0.0 (**6.6**$\downarrow$)	58.6 ± 0.5 (**3.8**$\downarrow$)	56.5 ± 0.4 (**7.4**$\downarrow$)
GAT-f†	60.8 ± 0.5 (**0.1**$\downarrow$)	58.2 ± 0.8 (**1.4**$\downarrow$)	56.5 ± 0.9 (**4.2**$\downarrow$)	57.2 ± 0.8 (**2.5**$\downarrow$)	57.6 ± 0.8 (**6.7**$\downarrow$)

Table A4. Attack success rate of two attacks with two order settings on ImageNet. The value in the parenthesis is the reduced value compare with another order settings.

F.2. How Do Composite Perturbations Fool the Model? Visual Examples

In Fig. A4, we present the inference results from an ℓ_∞-robust model (Madry$^\dagger_\infty$); the confidence bars are marked in green (red) if the prediction is correct (incorrect). The results showed that while a robust model can resist perturbations in ℓ_p ball, this only consideration is not comprehensive. That is, if we consider computing ℓ_∞ perturbations after some semantic attacks, the model may not exhibit the robustness it has around the ℓ_∞ ball.



(a) Attack order 1: $\underline{\ell_\infty}$ → Hue → Saturation → Rotation → Brightness → Contrast

(b) Attack order 2: Hue → Saturation → Rotation → $\underline{\ell_\infty}$ → Brightness→ Contrast

Figure A4. Warplane example. (a) The model can correctly predict objects under a given attack order. (b) The model can easily be fooled when adding ℓ_∞ perturbations after Hue, Saturation, and Rotation attacks. (*Note: Attack parameters are optimized by Comp-PGD algorithm.*)

G. Additional Experimental Results and Adversarial Examples

We further evaluate multiple CAAs in this section, and the experimental results on SVHN are also provided. In particular, we present the robust accuracy (RA) and their corresponding attack success rate (ASR). Again, the ASR is the percentage of the images that were initially classified correctly but were misclassified after being attacked; therefore, the lower ASR indicates the more robust model. In Sec. G.1, we especially show a single attack, which launches merely one attack from the attack pool. Notably, the ℓ_∞ (20-step) is regular PGD attack, and Auto-ℓ_∞ is an ensemble of four diverse attacks [1]. Multiple attacks (including three, semantic, and full) are listed in Sec. G.2. (For efficiency, we use ℓ_∞ (PGD) in multiple attack evaluation.)

G.1. Single Attack

Results on CIFAR-10

				Single attack				Auto attack
Training	Clean	Hue	Saturation	Rotation	Brightness	Contrast	ℓ_∞ (20-step)	Auto-ℓ_∞
Normal†	95.2	81.8 ± 0.0	94.0 ± 0.0	88.1 ± 0.1	92.1 ± 0.1	93.7 ± 0.1	0.0 ± 0.0	0.0 ± 0.0
Madry$^\dagger_\infty$	87.0	70.8 ± 0.0	84.8 ± 0.0	79.5 ± 0.1	77.0 ± 0.1	79.9 ± 0.1	53.5 ± 0.0	49.2 ± 0.0
PAT$^\dagger_{self}$	82.4	64.3 ± 0.2	79.8 ± 0.0	74.1 ± 0.1	72.5 ± 0.1	78.0 ± 0.1	41.2 ± 0.0	30.2 ± 0.0
PAT$^\dagger_{alex}$	71.6	53.2 ± 0.2	68.9 ± 0.0	63.8 ± 0.1	60.6 ± 0.1	65.2 ± 0.0	41.9 ± 0.0	28.8 ± 0.0
GAT-f†	82.3	81.2 ± 0.5	80.8 ± 0.1	78.3 ± 0.5	80.1 ± 0.1	79.7 ± 0.1	42.7 ± 0.0	38.7 ± 0.0
GAT-fs†	82.1	80.6 ± 0.0	80.8 ± 0.0	78.0 ± 0.2	80.4 ± 0.1	79.5 ± 0.1	46.6 ± 0.0	41.9 ± 0.0
Normal*	94.0	75.8 ± 0.2	92.3 ± 0.0	87.4 ± 0.2	89.1 ± 0.0	91.3 ± 0.1	0.0 ± 0.0	0.0 ± 0.0
Trades$^*_\infty$	84.9	65.7 ± 0.2	82.7 ± 0.0	77.5 ± 0.1	69.7 ± 0.3	70.7 ± 0.1	55.8 ± 0.0	52.5 ± 0.0
FAT$^*_\infty$	88.1	69.0 ± 0.1	85.4 ± 0.0	77.0 ± 0.1	73.1 ± 0.4	76.5 ± 0.1	54.7 ± 0.0	51.5 ± 0.0
AWP$^*_\infty$	85.4	67.5 ± 0.1	83.0 ± 0.0	77.0 ± 0.2	68.3 ± 0.1	70.8 ± 0.0	59.4 ± 0.0	56.2 ± 0.0
GAT-f*	83.4	82.3 ± 0.6	81.8 ± 0.0	79.5 ± 0.4	81.7 ± 0.0	81.0 ± 0.1	43.6 ± 0.0	40.0 ± 0.0
GAT-fs*	83.2	81.5 ± 0.1	81.7 ± 0.0	78.8 ± 0.0	81.2 ± 0.0	80.7 ± 0.1	47.2 ± 0.0	42.2 ± 0.0

Table A5. Robust accuracy of single attack, which is one of semantic attacks, on CIFAR-10

				Single attack				Auto attack
Training	Clean	Hue	Saturation	Rotation	Brightness	Contrast	ℓ_∞ (20-step)	Auto-ℓ_∞
Normal†	0.0	14.4 ± 0.0	1.4 ± 0.0	8.0 ± 0.2	3.3 ± 0.1	1.6 ± 0.1	100.0 ± 0.0	100.0 ± 0.0
Madry$^\dagger_\infty$	0.0	19.3 ± 0.0	2.6 ± 0.0	9.3 ± 0.1	11.7 ± 0.2	8.4 ± 0.0	38.6 ± 0.0	43.4 ± 0.0
PAT$^\dagger_{self}$	0.0	23.0 ± 0.2	3.3 ± 0.0	11.8 ± 0.2	12.6 ± 0.1	5.7 ± 0.1	50.0 ± 0.0	63.4 ± 0.0
PAT$^\dagger_{alex}$	0.0	27.7 ± 0.2	4.2 ± 0.1	13.0 ± 0.1	17.2 ± 0.2	10.0 ± 0.0	41.4 ± 0.0	59.8 ± 0.0
GAT-f†	0.0	1.6 ± 0.5	1.9 ± 0.1	5.7 ± 0.5	2.8 ± 0.1	3.3 ± 0.0	48.2 ± 0.0	53.0 ± 0.0
GAT-fs†	0.0	2.0 ± 0.0	1.6 ± 0.0	5.8 ± 0.2	2.2 ± 0.1	3.3 ± 0.1	43.2 ± 0.0	49.0 ± 0.0
Normal*	0.0	19.7 ± 0.2	1.8 ± 0.0	7.6 ± 0.2	5.3 ± 0.1	2.9 ± 0.1	100.0 ± 0.0	100.0 ± 0.0
Trades$^*_\infty$	0.0	23.0 ± 0.2	2.6 ± 0.0	9.5 ± 0.1	18.2 ± 0.3	16.9 ± 0.1	34.3 ± 0.0	38.2 ± 0.0
FAT$^*_\infty$	0.0	22.2 ± 0.1	3.1 ± 0.0	13.3 ± 0.1	17.3 ± 0.5	13.4 ± 0.1	37.9 ± 0.0	41.5 ± 0.0
AWP$^*_\infty$	0.0	21.5 ± 0.2	2.8 ± 0.0	10.5 ± 0.2	20.3 ± 0.1	17.3 ± 0.0	30.4 ± 0.0	34.2 ± 0.0
GAT-f*	0.0	1.7 ± 0.5	2.0 ± 0.0	5.6 ± 0.3	2.3 ± 0.0	3.1 ± 0.0	47.7 ± 0.0	52.0 ± 0.0
GAT-fs*	0.0	2.3 ± 0.1	1.9 ± 0.0	5.9 ± 0.0	2.6 ± 0.0	3.2 ± 0.0	43.3 ± 0.0	49.2 ± 0.0

Table A6. Attack success rate of single attack on CIFAR-10.

Results on ImageNet

				Single attack				Auto attack
Training	Clean	Hue	Saturation	Rotation	Brightness	Contrast	ℓ_∞ (20-step)	Auto-ℓ_∞
Normal†	76.1	50.9 ± 0.2	72.5 ± 0.1	68.2 ± 0.6	69.2 ± 0.3	71.8 ± 0.2	0.0 ± 0.0	0.0 ± 0.0
Madry$^\dagger_\infty$	62.4	38.1 ± 0.4	58.4 ± 0.0	51.0 ± 0.7	53.6 ± 0.1	55.9 ± 0.0	33.5 ± 0.0	28.9 ± 0.0
Fast-AT$^\dagger_\infty$	53.8	27.8 ± 0.2	48.0 ± 0.0	38.6 ± 0.6	42.0 ± 0.1	44.0 ± 0.0	27.5 ± 0.0	24.7 ± 0.0
GAT-f†	60.0	51.0 ± 2.5	58.1 ± 0.0	56.5 ± 0.3	57.7 ± 0.1	58.1 ± 0.1	25.2 ± 0.0	20.9 ± 0.0

Table A7. Robust accuracy of single attack, which is one of semantic attacks, on ImageNet.

				Single attack				Auto attack
Training	Clean	Hue	Saturation	Rotation	Brightness	Contrast	ℓ_∞ (20-step)	Auto-ℓ_∞
Normal†	0.0	34.3 ± 0.2	4.9 ± 0.1	12.2 ± 0.6	9.5 ± 0.3	5.9 ± 0.2	100.0 ± 0.0	100.0 ± 0.0
Madry$^\dagger_\infty$	0.0	39.9 ± 0.6	6.7 ± 0.0	19.9 ± 0.9	14.6 ± 0.1	10.8 ± 0.1	46.3 ± 0.0	53.6 ± 0.0
Fast-AT$^\dagger_\infty$	0.0	49.3 ± 0.3	11.1 ± 0.0	29.8 ± 0.9	22.8 ± 0.1	19.0 ± 0.1	48.9 ± 0.0	54.1 ± 0.0
GAT-f†	0.0	17.7 ± 3.2	3.3 ± 0.0	6.8 ± 0.3	3.9 ± 0.1	3.3 ± 0.1	57.9 ± 0.0	65.1 ± 0.0

Table A8. Attack success rate of single attack, on ImageNet

Results on SVHN

				Single attack				Auto attack
Training	Clean	Hue	Saturation	Rotation	Brightness	Contrast	ℓ_∞ (20-step)	Auto-ℓ_∞
Normal*	95.4	93.3 ± 0.0	94.7 ± 0.0	89.7 ± 0.1	92.2 ± 0.0	93.7 ± 0.0	0.5 ± 0.0	0.0 ± 0.0
Trades$^*_\infty$	90.3	87.3 ± 0.1	89.2 ± 0.0	81.4 ± 0.0	77.2 ± 0.1	83.3 ± 0.1	53.3 ± 0.0	44.2 ± 0.0
GAT-f*	93.4	92.5 ± 0.0	93.0 ± 0.0	91.2 ± 0.0	92.1 ± 0.1	92.1 ± 0.0	51.2 ± 0.0	36.9 ± 0.0
GAT-fs*	93.6	92.8 ± 0.0	93.1 ± 0.0	91.7 ± 0.0	92.5 ± 0.0	92.3 ± 0.0	54.1 ± 0.0	38.2 ± 0.0

Table A9. Robust accuracy of single attack, which is one of semantic attacks, on SVHN.

				Single attack				Auto attack
Training	Clean	Hue	Saturation	Rotation	Brightness	Contrast	ℓ_∞ (20-step)	Auto-ℓ_∞
Normal*	0.0	2.4 ± 0.0	0.7 ± 0.0	6.2 ± 0.1	3.4 ± 0.0	1.8 ± 0.0	99.5 ± 0.0	100.0 ± 0.0
Trades$^*_\infty$	0.0	4.0 ± 0.1	1.3 ± 0.0	10.0 ± 0.0	14.9 ± 0.1	8.0 ± 0.0	41.0 ± 0.0	51.0 ± 0.0
GAT-f*	0.0	1.1 ± 0.0	0.5 ± 0.0	2.5 ± 0.0	1.5 ± 0.1	1.5 ± 0.0	45.2 ± 0.1	60.5 ± 0.0
GAT-fs*	0.0	1.0 ± 0.0	0.6 ± 0.0	2.2 ± 0.1	1.2 ± 0.0	1.4 ± 0.0	42.2 ± 0.0	59.1 ± 0.0

Table A10. Attack success rate of single attack, on SVHN

G.2. Multiple Attacks: Three attacks, Semantic attacks and Full attacks

We only provided the ASR of CIFAR-10 and ImageNet here; the RA can be found in Tables 1 and 2 of our paper. Again, the abbreviation used here is the same as in the paper.

Results on CIFAR-10

		Three attacks			Semantic attacks		Full attacks	
Training	Clean	CAA_{3a}	CAA_{3b}	CAA_{3c}	Rand.	Sched.	Rand.	Sched.
Normal†	0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	37.4 ± 0.3	53.6 ± 0.5	100.0 ± 0.0	100.0 ± 0.0
Madry$^{\dagger}_{\infty}$	0.0	64.6 ± 0.2	78.4 ± 0.6	78.0 ± 0.3	63.8 ± 0.2	75.5 ± 0.3	87.6 ± 0.2	95.8 ± 0.2
PAT$^{\dagger}_{self}$	0.0	74.7 ± 0.2	85.6 ± 0.5	78.3 ± 0.4	65.0 ± 0.3	78.8 ± 0.4	88.9 ± 0.4	96.9 ± 0.3
PAT$^{\dagger}_{alex}$	0.0	71.1 ± 0.5	82.5 ± 0.2	77.0 ± 0.6	67.6 ± 0.4	82.9 ± 0.6	85.7 ± 0.1	96.5 ± 0.3
GAT-f†	0.0	51.6 ± 0.2	59.6 ± 0.1	64.9 ± 0.3	15.2 ± 0.1	19.8 ± 0.1	63.5 ± 0.5	77.1 ± 0.4
GAT-fs†	0.0	47.0 ± 0.1	55.4 ± 0.2	60.5 ± 0.2	15.0 ± 0.2	18.8 ± 0.1	60.7 ± 1.0	73.5 ± 0.4
Normal*	0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	51.1 ± 0.4	68.3 ± 0.6	100.0 ± 0.0	100.0 ± 0.0
Trades$^{*}_{\infty}$	0.0	64.7 ± 0.4	76.7 ± 0.7	88.1 ± 0.6	80.4 ± 0.3	90.5 ± 0.5	93.2 ± 0.4	98.2 ± 0.2
FAT$^{*}_{\infty}$	0.0	66.2 ± 0.4	80.6 ± 0.5	85.5 ± 0.7	78.8 ± 0.2	88.9 ± 0.5	93.0 ± 0.2	98.3 ± 0.1
AWP$^{*}_{\infty}$	0.0	59.9 ± 0.2	72.8 ± 0.3	87.0 ± 0.4	81.8 ± 0.2	90.7 ± 0.2	93.1 ± 0.1	98.0 ± 0.2
GAT-f*	0.0	51.8 ± 0.1	59.2 ± 0.1	63.2 ± 0.5	14.2 ± 0.1	18.7 ± 0.2	62.6 ± 0.4	75.9 ± 0.4
GAT-fs*	0.0	47.8 ± 0.2	56.3 ± 0.1	60.5 ± 0.4	15.4 ± 0.1	19.8 ± 0.4	61.4 ± 0.9	73.7 ± 0.8

Table A11. Attack success rate of composite semantic attacks and composite full attacks on CIFAR-10.

Results on ImageNet

		Three attacks			Semantic attacks		Full attacks	
Training	Clean	CAA_{3a}	CAA_{3b}	CAA_{3c}	Rand.	Sched.	Rand.	Sched.
Normal†	0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	59.1 ± 0.5	72.9 ± 1.3	100.0 ± 0.0	100.0 ± 0.0
Madry$^{\dagger}_{\infty}$	0.0	77.7 ± 0.6	85.2 ± 0.4	74.0 ± 1.3	77.6 ± 0.1	85.6 ± 0.1	88.7 ± 0.2	95.5 ± 0.4
Fast-AT$^{\dagger}_{\infty}$	0.0	82.3 ± 0.5	89.8 ± 0.2	78.9 ± 1.5	88.3 ± 0.1	93.3 ± 0.1	94.3 ± 0.2	98.1 ± 0.2
GAT-f†	0.0	67.9 ± 1.7	68.4 ± 2.3	69.2 ± 0.7	27.8 ± 3.0	34.6 ± 3.3	69.1 ± 0.9	80.3 ± 0.2

Table A12. Attack success rate of composite semantic attacks and composite full attacks on ImageNet.

Results on SVHN

		Three attacks			Semantic attacks		Full attacks	
Training	Clean	CAA_{3a}	CAA_{3b}	CAA_{3c}	Rand.	Sched.	Rand.	Sched.
Normal*	95.4	0.4 ± 0.0	0.2 ± 0.0	0.4 ± 0.1	78.5 ± 0.2	68.7 ± 0.6	0.5 ± 0.0	0.2 ± 0.1
Trades$^{*}_{\infty}$	90.3	43.6 ± 0.1	32.1 ± 0.3	21.2 ± 0.7	47.3 ± 0.2	34.7 ± 0.5	22.6 ± 0.5	10.6 ± 0.4
GAT-f*	93.4	47.0 ± 0.1	42.8 ± 0.3	34.4 ± 0.5	85.5 ± 0.1	82.8 ± 0.2	37.1 ± 0.2	26.8 ± 0.6
GAT-fs*	93.6	48.7 ± 0.1	45.2 ± 0.3	35.6 ± 0.5	86.6 ± 0.1	83.7 ± 0.2	39.0 ± 0.4	28.2 ± 0.3

Table A13. Robust accuracy of composite semantic attacks and composite full attacks on SVHN.

Training	Clean	Three attacks			Semantic attacks		Full attacks	
		CAA_{3a}	CAA_{3b}	CAA_{3c}	Rand.	Sched.	Rand.	Sched.
Normal*	0.0	99.5 ± 0.0	99.8 ± 0.0	99.6 ± 0.1	17.8 ± 0.3	28.0 ± 0.6	99.5 ± 0.0	99.8 ± 0.1
Trades$_\infty^*$	0.0	51.7 ± 0.1	64.4 ± 0.3	76.5 ± 0.8	47.6 ± 0.3	61.6 ± 0.6	75.0 ± 0.5	88.2 ± 0.5
GAT-f*	0.0	49.7 ± 0.1	54.2 ± 0.3	63.2 ± 0.5	8.5 ± 0.1	11.4 ± 0.2	60.3 ± 0.2	71.3 ± 0.7
GAT-fs*	0.0	48.0 ± 0.1	51.7 ± 0.3	62.0 ± 0.6	7.5 ± 0.1	10.6 ± 0.2	58.3 ± 0.4	69.8 ± 0.3

Table A14. Attack success rate of composite semantic attacks and composite full attacks on SVHN.

H. Examples of Single Semantic Attacks at Different Levels

Fig. A5 shows five single semantic attacks with corresponding perturbation levels. Each row represents the perturbed image of a corresponding attack A_k with different perturbation values $\delta_k \in \epsilon_k$.

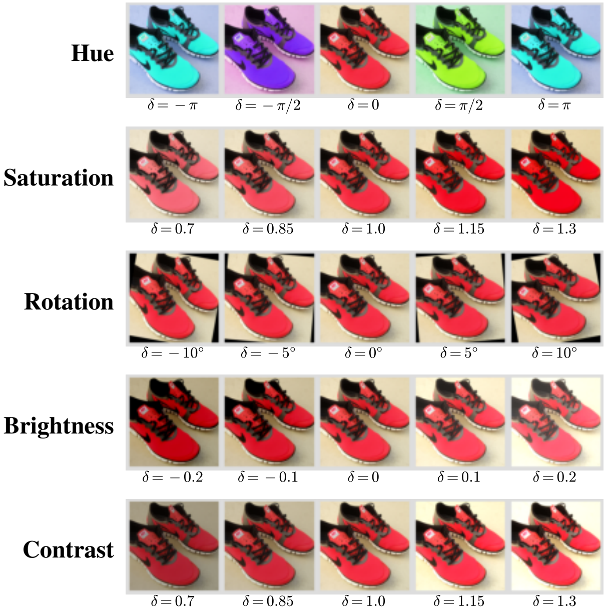


Figure A5. Single Semantic Attack Examples. Clean image was placed at the center of each row.

I. Additional Visualization of Adversarial Examples under Different CAA

We present a series of adversarial examples from ImageNet generated using our proposed composite adversarial attacks (CAA). These attacks include single attacks, two attacks, three attacks, semantic attacks, and full attacks. To illustrate the effectiveness of our approach, we provide visualizations of the adversarial examples for each type of attack, arranged in several columns. Specifically, the left-most column of Figures A6, A7, and A8 shows the original images. Every of the following two columns are the adversarial examples generated from one of the CAA attacks and their differences compared with the original images. Note that for visualization purposes only, all differences have been multiplied by three.

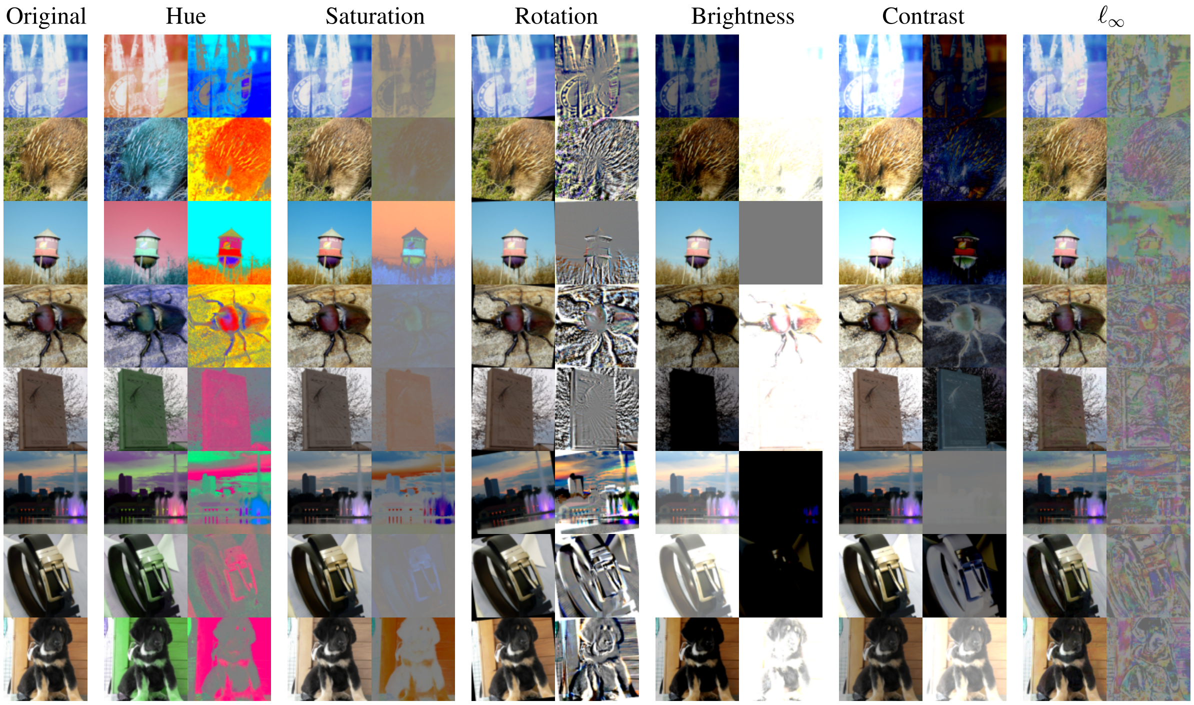


Figure A6. Adversarial examples generated under **single semantic** attacks or ℓ_∞ attack.

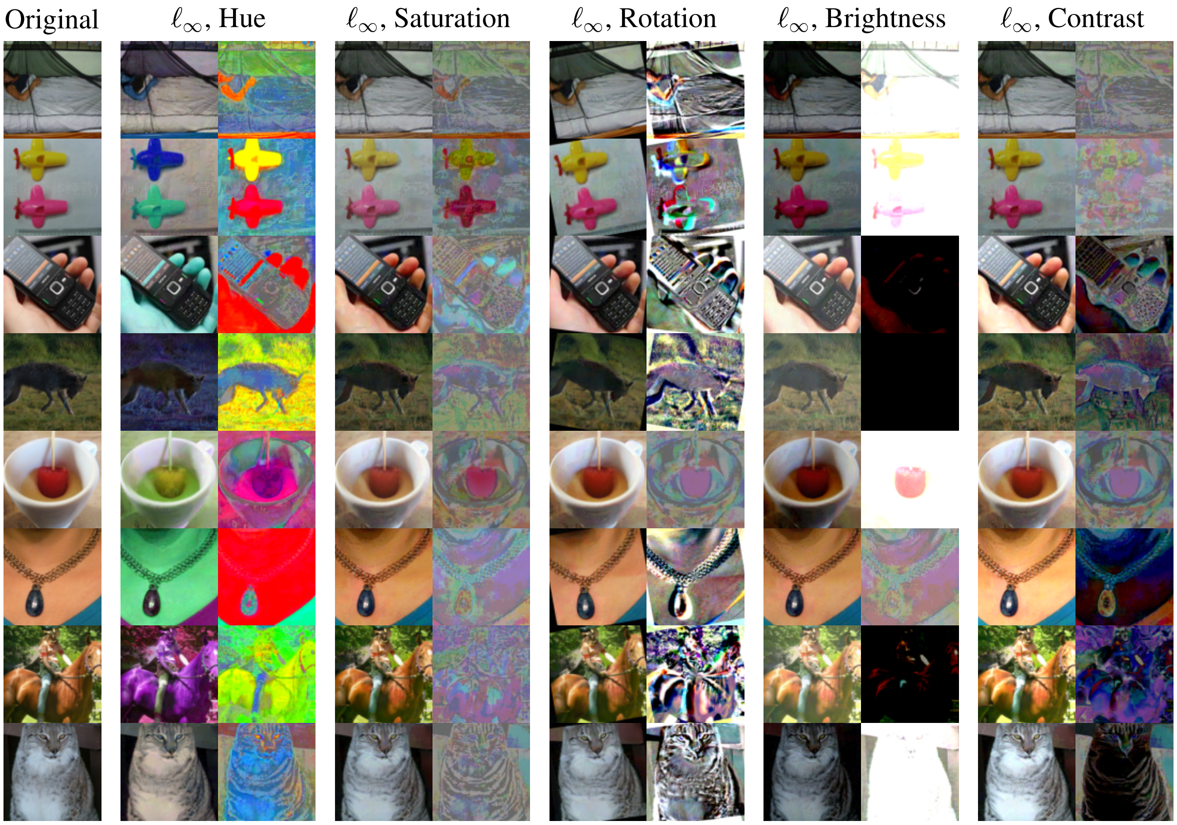


Figure A7. Composite adversarial examples generated under **two attacks** (composed of one semantic attack and the ℓ_∞ attack).

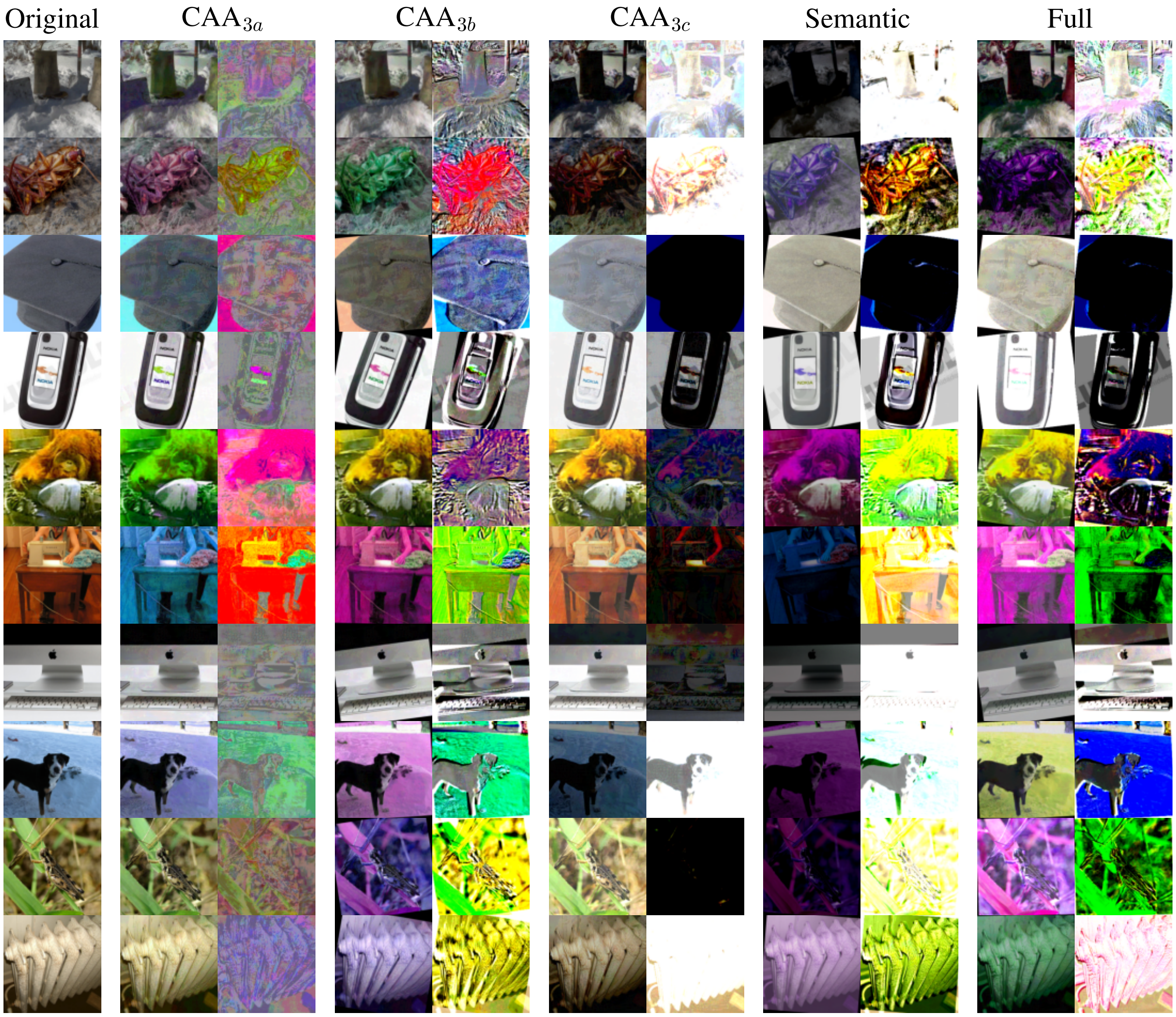


Figure A8. Composite adversarial examples generated under **three attacks and other multiple attacks**. *CAA*$_{3a}$: (Hue, Saturation, ℓ_∞), *CAA*$_{3b}$: (Hue, Rotation, ℓ_∞), *CAA*$_{3c}$: (Brightness, Contrast, ℓ_∞), Semantic: (Hue, Saturation, Rotation, Brightness, and Contrast), and Full: ℓ_∞+ *all semantic attacks.*

References

[1] 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. 6

[2] 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. 1, 2

[3] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. 1

[4] 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

[5] 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. 1

[6] 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. 1, 2

[7] 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, 4