Beyond ℓ_p Tunnel Vision: An Exploration of Robustness to Multiple Perturbation Types and their Compositions

Luke Rowe[*]
Department of Computer Science
University of Waterloo
l6rowe@uwaterloo.ca

Benjamin Thérien[*]
Department of Computer Science
University of Waterloo
btherien@uwaterloo.ca

Abstract

In adversarial machine learning, the popular ℓ_p threat model has been the focus of much previous work [21, 11]. While this mathematical definition of imperceptibility successfully captures an infinite set of additive image transformations that a model should be robust to, this is only a subset of all transformations which leave the semantic label of an image unchanged. Although previous work advocates for the consideration of spatial attacks [7], in doing so, they sidestep the ℓ_p threat model, failing to consider both together. Intuitively, the composition of these two transformations should lead to an imperceptibly transformed image under a strictly stronger threat model, yet little prior work considers this setting. In the following, we highlight shortcomings of the longstanding ℓ_p threat model and propose extending it to a more realistic setting which composes ℓ_p perturbations with spatial transformations. As a first investigation, we adapt state of the art ℓ_p defenses to this novel threat model and study their performance against compositional attacks. We find that our newly proposed $\text{TRADES}_{\text{All}}$ strategy performs strongest of all.

1 Introduction

Despite the outstanding performance of deep neural networks[25, 6, 26] on a variety of computer vision tasks, deep neural networks have been shown to be vulnerable to human-imperceptible adversarial perturbations [21, 11]. Designing algorithms to train models that are robust to small human-imperceptible ℓ_p-bounded alterations of the input has been an extensive focus of previous work [16, 27]. While it is certainly unreasonable for a classifier to change its decision based on the addition of imperceptible ℓ_p-bounded noise, this is not the only input transformation we wish to be robust to. Many spatial transformations leave an image's label unchanged, but are ill-defined by an ℓ_p-threat model (see fig. 1). Yet, any classifier deemed robust should not be any more vulnerable to the affine transformed images seen in fig. 1, simply because these are ill-defined in the ℓ_p-threat model. To further demonstrate this, we plot affine transformed images for the entire CIFAR-10 dataset, showing that no pair of natural and affine transformed images is considered valid under the ℓ_p threat model, despite being very perceptually similar to a human and certainly semantically equivalent. This highlights the need to consider threat models which go beyond the ℓ_p setting, allowing for the consideration of such images.

While certainly many prior works have considered robustness under adversarial settings that differ from the standard ℓ_p setting, these existing works either consider robustness under a single perturbation type [24, 1, 9, 14, 8, 13] or by selecting from a fixed set (*i.e.,* the union) of perturbation types [17, 18, 22]. However, relatively few consider robustness to the composition of multiple perturbation types [23]. Realistically, an adversary is not restricted to selecting a perturbation from one threat

Preprint. Under review.

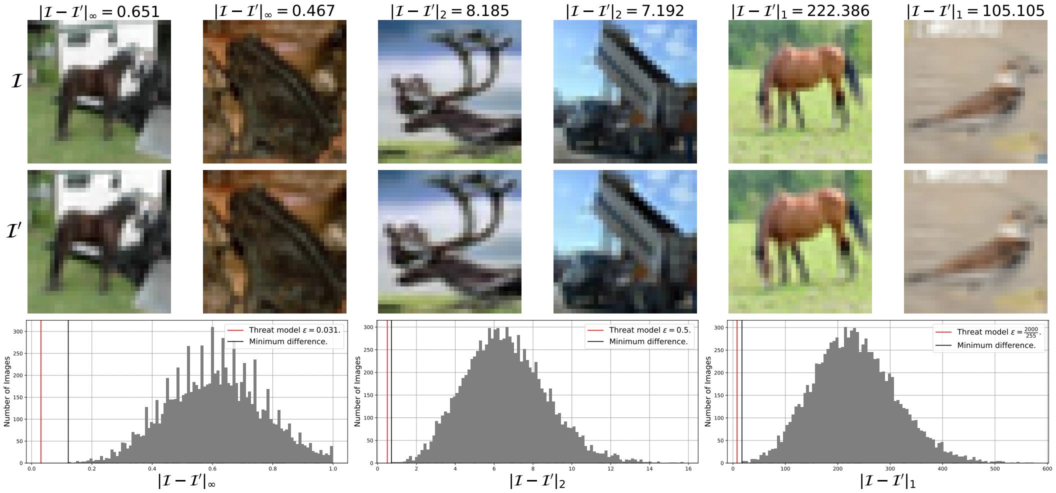

Figure 1: ℓ_p **norm threat models do not include other transformations of interest.** The first two rows show pairs of clean images ($\mathcal{I}$) from the CIFAR-10 test set and these same images scaled, rotated, and translated using bilinear interpolation ($\mathcal{I}'$). The three plots show the frequency distribution of $\|\mathcal{I} - \mathcal{I}'\|_p$ for $p \in \{1, 2, \infty\}$.

model but may choose to compose perturbations from multiple threat models (see fig. 2). Moreover, recent work has shown that under a simple statistical setting, defending against an adversary who can compose ϵ-bounded ℓ_∞ perturbations and rotation/translation (RT) transformations is strictly harder than defending against an adversary who can select from the union [22]. This theoretical result highlights the need to explore how we can build truly robust models in this well-motivated compositional setting. As a first step toward addressing this prominent gap in the adversarial robustness literature, we aim to broaden the notion of human-imperceptible transformations by specifically considering the composition of bounded ℓ_∞ perturbations and RT transformations of the input. Our main contributions are three-fold:

- We demonstrate the need for truly robust models in the compositional setting.
- We train a family of empirical defenses constructed from TRADES [27] under various threat models – including ℓ_∞, RT, and their union and composition.
- We benchmark our family of TRADES-trained models on a suite of competitive white-box attacks and provide thorough analysis of our results and implications for future work.

The remainder of the paper is organized as follows: Section 2 overviews related prior work and highlights our novel contributions to this body of literature, Section 3 details the family of TRADES defense models we consider and corresponding attack strategies, Section 4 specifies training hyperparameters and explains the results of our empirical evaluation, Section 5 discusses the most important implications of our findings, limitations, and possible directions of future work, and Section 6 overviews the general findings of our work.

2 Related work

Most existing works that consider threat models beyond the ℓ_p setting consider spatial transformations of the input [24, 1, 9, 14, 8, 13]. Engstrom et al. showed that the simple spatial threat model consisting of bounded rotations and translations was sufficient to significantly degrade the accuracy of standard image classifiers, even when the models are trained with spatial data augmentation [9, 13]. Concurrently, Xiao et al. showed that image classifiers are vulnerable to flow-induced spatial perturbations of the input image [24]. Similar to the ℓ_p setting, defenses against such spatial attacks can be broadly grouped into two categories: empirical defenses largely based on adversarial training [8] and certified defenses [1, 14].

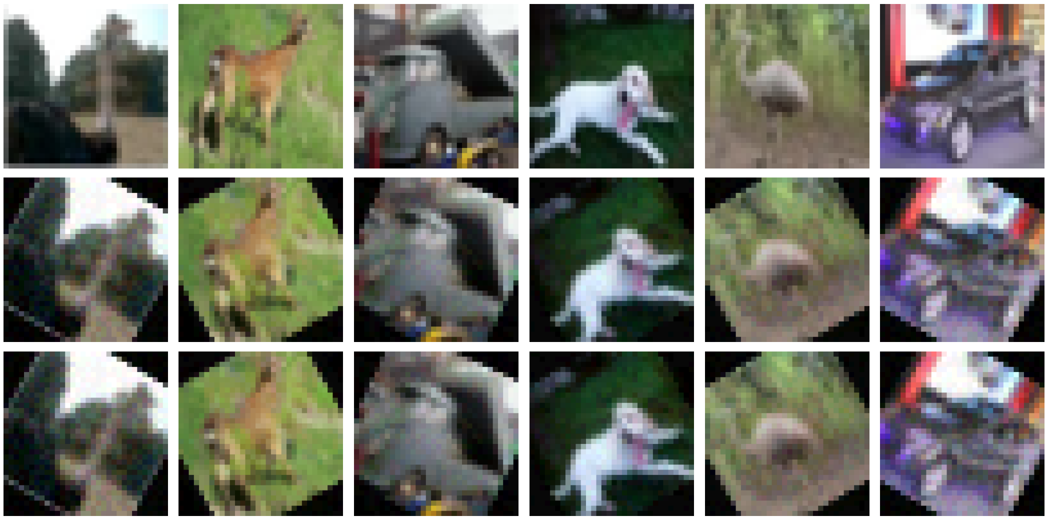

Figure 2: **Adversarial images obtained by the AAA $\circ$ RT attack on the TRADES$_{\text{All}}$ model**. The first row shows clean images. The second row shows the same images with applied adversarial RT transformations, where rotation angle $|\theta| \leq 30$ and horizontal/vertical translations $|\delta_x| \leq 3$px, $|\delta_y| \leq 3$px. The third row show the same RT images as above, but with imperceptible ℓ_∞ noise under the $\| \cdot \|_\infty \leq 0.031$ threat model.

Although many existing works consider robustness to spatial transformations, relatively few works consider the problem of jointly attaining robustness to both spatial transformations and ℓ_p perturbations of the input [17, 18, 22]. Existing works in the space of multiple perturbation robustness can be broadly grouped into two categories: adversarial training methods [18, 22], which focus on extending the adversarial reachable set to cover the union of multiple threat models; and perturbation type classification methods [17], which are methods to identify the perturbation type being applied by the attacker, enabling intelligent input routing to appropriately trained robust networks. Furthermore, recent works have reported the existence of an ℓ_p-spatial robustness tradeoff, both theoretically and empirically on popular image datasets [18, 12]. These existing works on multiple perturbation robustness, however, mainly focus on robustness to the union of multiple perturbations models, and little prior work considers the compositional setting [23]. We later discovered a paper that performs similar experiments to ours in this compositional setting; however, they perform standard adversarial training [16], whereas we construct robust models in this compositional setting from TRADES [27].

3 Methodology

3.1 The Compositional Threat Model

In this work, we consider a compositional threat model consisting of the composition of ϵ-bounded ℓ_∞ perturbations and bounded RT transformations. Concretely, for the ℓ_∞ threat model, we consider an adversary who can perturb an image $\boldsymbol{X}$ by adding ϵ-bounded ℓ_∞ noise to $\boldsymbol{X}$. That is, the adversarial reachable region $\mathcal{A}^{\ell_\infty}(\boldsymbol{X})$ under the ℓ_∞-threat model is defined by:

$$\mathcal{A}^{\ell_\infty}(\boldsymbol{X}) = \mathbb{B}_\infty(\boldsymbol{X}, \epsilon) := \{\boldsymbol{X} + \boldsymbol{\Delta}; ||\boldsymbol{\Delta}||_\infty \leq \epsilon\}. \tag{1}$$

For the RT threat model, we consider an adversary who can apply a bounded rotation θ followed by a bounded horizontal and vertical translation δ_x, δ_y to $\boldsymbol{X}$. Concretely, the adversarial reachable region under the RT-threat model is defined by:

$$\mathcal{A}^{\text{RT}}(\boldsymbol{X}) = \{\mathcal{T}(\boldsymbol{X}; \theta, \delta_x, \delta_y); |\theta| \leq \theta^{\max}, |\delta_x| \leq \delta_x^{\max}, |\delta_y| \leq \delta_y^{\max}\}, \tag{2}$$

where $\mathcal{T}(\cdot\,; \theta, \delta_x, \delta_y)$ is the affine transformation function with rotation θ and horizontal/vertical translations δ_x, δ_y, which implicitly warps the image via an interpolation algorithm (our experiments utilize bilinear interpolation). For the compositional threat model, the adversarial reachable region is naturally defined by:

$$\mathcal{A}^{\ell_\infty \circ \text{RT}}(\boldsymbol{X}) = \{\mathbb{B}(\mathcal{T}(\boldsymbol{X}; \theta, \delta_x, \delta_y), \epsilon); |\theta| \leq \theta^{\max}, |\delta_x| \leq \delta_x^{\max}, |\delta_y| \leq \delta_y^{\max}\}. \tag{3}$$

That is, $\mathcal{A}^{\ell_\infty \circ \text{RT}}(\boldsymbol{X})$ is defined as as the set of ϵ-bounded ℓ_∞ balls around all valid affine transformations of the image $\boldsymbol{X}$. It is important to note that prior work that considers the compositional threat model [22] considers convex combinations of ℓ_∞-perturbations and RT-transformations. However, as these perturbation models are fairly orthogonal, we believe that convex combinations of such transformations unreasonably limit the strength of a compositional adversary. Thus, our compositional threat model is a simple addition of both perturbation types, rather than a convex combination. To compare with the compositional setting, we also consider the union threat model consisting of the union of ϵ-bounded ℓ_∞ perturbations and bounded RT transformations [22]. In this case, the adversarial reachable region under the $\ell_\infty \cup$ RT-threat model is defined as the union of the adversarial reachable regions for the ℓ_∞-threat model and RT-threat model:

$$\mathcal{A}^{\ell_\infty \cup \text{RT}}(\boldsymbol{X}) = \mathbb{B}(\boldsymbol{X}, \epsilon) \cup \{\mathcal{T}(\boldsymbol{X}; \theta, \delta_x, \delta_y); |\theta| \leq \theta^{\max}, |\delta_x| \leq \delta_x^{\max}, |\delta_y| \leq \delta_y^{\max}\}. \tag{4}$$

3.2 Proposed defense methods

To explore the space of compositional adversarial examples and the compositional threat model, we train a family of empirical defenses constructed from TRADES [27], and evaluate these defenses in a white-box setting. We choose a white-box setting so that we can assess the full adversarial strength of these compositional adversarial examples. Below, we have a general form for the TRADES objective:

$$\min_f \mathbb{E}\Bigg\{ \underbrace{\mathcal{L}(f(\boldsymbol{X}), Y)}_{\text{Natural Accuracy}} + \underbrace{\max_{\boldsymbol{X}' \in \mathcal{A}(\boldsymbol{X})} \mathcal{L}(f(\boldsymbol{X}), f(\boldsymbol{X}'))/\lambda}_{\text{Robustness under } \mathcal{A}(\boldsymbol{X})} \Bigg\}. \tag{5}$$

The TRADES training objective augments the standard ERM objective with a robust regularization term that encourages robustness to examples within the defined adversarial reachable region $\mathcal{A}(\boldsymbol{X})$ by imposing a cross-entropy loss between the logits of the natural image $\boldsymbol{X}$ and the logits of the worst-case adversarial image $\boldsymbol{X'}$ within $\mathcal{A}(\boldsymbol{X})$. The hyperparameter λ balances the tradeoff between adversarial robustness and natural accuracy; a lower value of λ leads to a more adversarially robust model, but typically at the cost of reduced natural accuracy. We train a family of TRADES models under the various threat models discussed in Section 3.1. Concretely, we train the following family of TRADES defense methods:

- $\text{TRADES}_{\ell_\infty}$
- $\text{TRADES}_{\text{RT}}$
- $\text{TRADES}_{\ell_\infty \cup \text{RT}}$
- $\text{TRADES}_{\ell_\infty \circ \text{RT}}$

The subscript indicates the threat model being considered during training. To train TRADES models under these new threat models, we require a way to efficiently solve the inner optimization problem in the TRADES objective for the new corresponding definitions of $\mathcal{A}(\boldsymbol{X})$. In the ℓ_∞ case, a popular strategy is to perform Projected Gradient Descent (PGD) for a small number of steps, and we follow this strategy as well. For the RT-threat model, we perform **Worst-of-10** search: we simply take 10 random valid affine transformations and select the affine transformed image that attains the highest loss. This method is used in prior work [8]. For the loss function, we use the KL-divergence between the logits of the natural image and the logits of the transformed image. This method may seem very crude, but it works well in practice as we are only optimizing 3 parameters in the RT case – namely, the rotation value θ and the horizontal and vertical translation values δ_x, δ_y. In fact, recent work has shown that the loss landscape for adversarial training of spatial transformations is highly non-convex, so Worst-of-10 search outperforms first-order methods for solving the inner maximization problem in this case [8].

For the union setting, we use an existing competitive approach called the **Max Strategy** [22], in which we compute an ℓ_∞ perturbation using PGD and an RT perturbation using Worst-of-10, and select the perturbation that attains the maximum loss. Here, again, we use the KL-divergence for the loss function. For the compositional setting, we propose the **Worst-of-Worst** strategy, whereby

Defense / Attack	β	AAA $\circ$ RT	PGD $\circ$ RT	AAA $\cup$ RT	PGD $\cup$ RT	AAA	PGD	RT	Natural
TRADES$_{\text{All}}$	3.0	37.97	42.28	48.78	52.61	48.51	52.66	81.79	85.49
TRADES$_{\ell_\infty \circ \text{RT}}$	3.0	31.85	36.47	40.30	44.62	40.32	44.66	78.47	83.81
TRADES$_{\ell_\infty \cup \text{RT}}$	3.0	9.67	15.10	47.69	51.22	47.92	51.48	84.69	86.42
TRADES$_{\ell_\infty}$	3.0	5.27	6.32	18.73	19.70	51.19	54.14	29.18	86.36
TRADES$_{\text{RT}}$	3.0	0.00	0.00	0.00	0.00	0.00	0.00	**89.33**	**95.46**
TRADES$_{\text{All}}$	6.0	**39.82**	**44.44**	49.45	53.94	49.45	53.97	79.76	83.65
TRADES$_{\ell_\infty \circ \text{RT}}$	6.0	25.62	30.46	39.63	44.14	39.86	44.47	74.76	82.77
TRADES$_{\ell_\infty \cup \text{RT}}$	6.0	15.94	24.15	**50.56**	**54.64**	50.65	54.77	83.67	84.72
†TRADES$_{\ell_\infty}$	6.0	7.24	9.11	20.77	21.93	**53.02**	**56.63**	32.36	84.92
TRADES$_{\text{RT}}$	6.0	0.00	0.00	0.00	0.00	0.00	0.01	89.05	94.91
Natural	-	0.00	0.00	0.00	0.00	0.00	0.00	40.96	93.68

† from [27].

Table 1: **CIFAR-10 results table for different defense methods.** Columns correspond to accuracy under different perturbation types, while rows correspond to different defense models. All RT attacks utilize our grid search strategy. PGD attacks use 20 iterations on CIFAR-10. The best performing entry under a given attack is **bolded**, while the second best is underlined.

we first compute an RT adversarial example using Worst-of-10, and then we perform PGD on the RT-perturbed image. It is important to note that Worst-of-Worst implicitly assumes that the "worst" adversarial image from Worst-of-10 will produce the "worst" compositional adversarial example. We leave it to future work to verify whether this assumption holds in practice.

3.2.1 TRADES$_{\text{All}}$

We observe that the family of TRADES models proposed in Section 3.2 each train exclusively on adversarial images tailored to their respective threat models. However, this may not strike a favourable balance in performance between the different threat models at evaluation time. To address this issue, we propose TRADES$_{\text{All}}$, whereby adversarial training alternates between ℓ_∞ adversarial examples, RT adversarial examples and compositional adversarial examples. Concretely, given training image $\boldsymbol{X}$, TRADES$_{\text{All}}$ selects uniformly at random between a corresponding ℓ_∞ adversarial example, RT adversarial example and a compositional adversarial example for $\boldsymbol{X}$ when solving the inner maximization problem. The aim of this defense is to strike the right balance between all these perturbation types, without focusing too much on a single perturbation type. Note that TRADES$_{\text{All}}$ is still under the compositional threat model, as compositional adversarial examples are used for adversarial training $1/3$ of the time.

3.3 Attack methods background

We now describe the white-box attack algorithms used to evaluate our family of TRADES defenses. For the ℓ_∞-based attacks, we evaluate on Adaptive AutoAttack [15], or AAA, which is a recently published state-of-the-art adaptive white box attack. At a high level, AAA can be viewed as a variant of PGD with two major improvements: AAA chooses a better perturbation initialization strategy (recall that PGD initializes the perturbation randomly), and AAA is adaptive in that it allocates more of its attack budget towards samples that are easy to attack. The exact algorithmic details of AAA can be found at [15]. For the RT based-attack, we perform a simple grid-search on the 3 parameters $\theta, \delta_x, \delta_y$ that define the affine transformation. The grid search involves evenly-spaced values for each parameter: θ has 12 values, and δ_x, δ_y each have 5 values. Prior work has shown that grid search performs state-of-the-art for RT-based attacks [8].

Our complete attack suite is listed below:

- AAA
- PGD
- RT (Grid Search)
- AAA $\cup$ RT
- PGD $\cup$ RT
- AAA $\circ$ RT

Defense / Attack	β	AAA $\circ$ RT	PGD $\circ$ RT	AAA $\cup$ RT	PGD $\cup$ RT	AAA	PGD	RT	Natural
TRADES$_{\text{All}}$	1.0	59.99	79.20	92.26	95.16	92.51	95.60	97.61	99.34
TRADES$_{\ell_\infty \circ \text{RT}}$	1.0	65.76	81.78	91.37	94.51	91.64	95.10	97.28	99.03
TRADES$_{\ell_\infty \cup \text{RT}}$	1.0	36.23	66.31	91.30	94.89	91.37	95.20	97.85	99.47
TRADES$_{\ell_\infty}$	1.0	16.22	33.64	74.11	75.58	92.99	95.88	76.63	99.52
TRADES$_{\text{RT}}$	1.0	0.00	0.09	0.00	0.18	0.00	0.18	99.02	**99.64**
TRADES$_{\text{All}}$	3.0	65.36	80.97	92.11	94.56	92.30	95.00	97.36	98.94
TRADES$_{\ell_\infty \circ \text{RT}}$	3.0	**68.70**	83.08	90.87	93.68	91.03	94.01	96.96	98.65
TRADES$_{\ell_\infty \cup \text{RT}}$	3.0	45.31	73.78	92.12	**95.62**	92.20	95.62	97.78	99.28
TRADES$_{\ell_\infty}$	3.0	19.29	38.23	74.13	75.25	**93.83**	**96.48**	76.20	99.35
TRADES$_{\text{RT}}$	3.0	0.00	0.00	0.00	0.17	0.00	0.17	**99.13**	99.49
TRADES$_{\text{All}}$	6.0	67.57	**83.56**	91.89	94.76	92.14	95.25	97.05	98.73
TRADES$_{\ell_\infty \circ \text{RT}}$	6.0	68.52	82.66	90.39	93.29	90.61	93.75	96.70	98.22
TRADES$_{\ell_\infty \cup \text{RT}}$	6.0	48.12	76.43	**92.37**	95.09	92.52	95.46	97.61	99.22
†TRADES$_{\ell_\infty}$	6.0	14.66	36.45	73.71	75.52	92.68	96.07	76.51	99.48
TRADES$_{\text{RT}}$	6.0	0.00	0.00	0.00	0.00	0.00	0.00	98.79	99.42
Natural	-	0.00	0.14	0.00	2.13	0.00	2.18	67.82	99.18

† from [27].

Table 2: **MNIST results table for different defense methods.** Columns correspond to accuracy under different perturbation types, while rows correspond to different defense models. All RT attacks utilize our grid search strategy. PGD attacks use 40 iterations on MNIST. The best performing entry under a given attack is **bolded**, while the second best is underlined.

- PGD $\circ$ RT

The multiple perturbation attacks (*i.e.*, the union and compositional attacks) operate in the same manner as in solving the inner maximization problem during TRADES training. For example, the union attacks use the Max Strategy and the composition attacks use the Worst-of-Worst strategy.

4 Empirical Evaluation & Results

In this section, we present our empirical evaluation of the proposed family of TRADES defense models trained and evaluated on MNIST and CIFAR-10 in the threat models introduced above.

4.1 Training setup

We take a number of steps to improve the reproducibility of our results and comparability to prior work. Firstly, to train our family of TRADES defenses we utilize the default hyperparameters included in the author's GitHub repository[1]. Secondly, we use the same seed for each experiment to obtain the same random weight initializations and dataset shuffles. Thirdly, to attack our defense models, we use the default code provided by Liu et al. [15] for their AAA implementation and we borrow more code from the TRADES repository for PGD attacks, where PGD attacks on CIFAR-10 are run for 20 iterations with $\epsilon = 0.031$, while attacks on MNIST are run for 40 iterations with $\epsilon = 0.3$ (the same ϵ values are used for AAA). Lastly, in the RT-threat model, we set $\theta^{\max} = 30$, $\delta_x^{\max} = 3\text{px}$, $\delta_y^{\max} = 3\text{px}$, which is consistent with prior work [8].

4.2 Results

The results of our empirical evaluations are reported in Tables 1 and 2. We note that the AAA attack performs stronger than PGD on all experiments. Therefore, we consider PGD as a point of comparison to prior work, but focus on attacks generated by AAA to measure robust accuracy and the relative performance of our models. Many general trends can be observed across both datasets, while the best performing models differ from dataset to dataset and other small differences are observed.

On both datasets, we observe that composition attacks are strictly empirically stronger than union attacks. This finding aligns with results from [22] that show the inequality to be strict in a simple statistical setting. Furthermore, ℓ_∞ attacks are strictly empirically stronger than their RT counterparts against all non-specialized models. Analyzing the effect of β, we note that in general ℓ_∞ robustness

[1] see "train_trades_cifar10.py" and "train_trades_mnist.py" at `https://github.com/yaodongyu/TRADES`

Training algo.	Natural	$\text{TRADES}_{\ell_\infty}$	$\text{TRADES}_{\text{All}}$	$\text{TRADES}_{\text{RT}}$	$\text{TRADES}_{\ell_\infty \circ \text{RT}}$	$\text{TRADES}_{\ell_\infty \cup \text{RT}}$
Training Time MNIST	$00:16$	$01:52$	$05:49$	$07:05$	$08:24$	$08:32$
Training Time CIFAR-10	$03:09$	$35:25$	$30:21$	$21:12$	$47:28$	$49:36$

Table 3: **Training times in HH:MM format for different defense models.** All models were trained on a single NVIDIA V100 GPU with the same hyperparamters. The only difference between each entry is their way of solving the inner-maximization problem, decribed in sec.3.2. We note that reported times include one AAA attack performed after training which introduces some variance of approximately ±30 mins for CIFAR-10 and ±7 mins for MNIST.

benefits from higher values, while robustness to RT attacks decreases as β is increased. This seems to suggest that stronger representation stability tending toward invariance should be encouraged for ℓ_p noise, while more freedom is desirable for learning representations stable to rotations and translations which generalize to other settings. Finally, we note that models trained on RT exhibit greater robustness to this attack than ℓ_∞ trained models do on ℓ_∞ attacks, showing that ℓ_∞ is harder to defend against.

Table 1 reports the robust and natural accuracy of TRADES defense models trained on CIFAR-10 for $\beta \in \{3.0, 6.0\}$ and one naturally trained model of the same architecture. The sixth row of the table showcases the strong overall performance of the All strategy at $\beta = 6$, which performs best on composition attacks and when accounting for all settings together. Its improved performance against composition attacks when compared to the model trained exclusively on compositions suggests that alternating training schemes can be beneficial in this setting. Moreover, the mediocre performance of the union trained models in these settings demonstrates that training on the union is insufficient to defend against a composite adversary. The RT attacks are best prevented by the $\text{TRADES}_{\text{RT}}$ models, as expected, shortly followed by the union models, and the All model. Finally, these RT models also perform best on natural accuracy, even when compared to the naturally trained classifier, though this may be attributable to the rudimentary data augmentation used by [27] in their repository which may lead the natural model to underperform.

Table 2 reports the robust and natural accuracy of TRADES defense models trained on MNIST for $\beta \in \{1.0, 3.0, 6.0\}$ and one naturally trained model of the same architecture. We observe that unlike for CIFAR-10, where the All-trained models shines against the composition attack, for MNIST specialized models always outperform their more general counterparts except for natural accuracy where $\text{TRADES}_{\text{RT}}$ at $\beta = 1.0$ performs best. However, the All models trained at the same β value are still competitive and lose at most 2 percentage points of accuracy when compared to the best performing model on each attacks. When comparing among different β values for the $\text{TRADES}_{\text{All}}$ model, we observe saturating performance gains against composition attacks as β is increased. For the RT-trained models, we observe that the accuracy of RT attacks generally decreases as β is increased, which is consistent with our CIFAR-10 results. In contrast to CIFAR-10 results, however, ℓ_∞ robustness does not seem to increase linearly with β, except for the $\text{TRADES}_{\ell_\infty \cup \text{RT}}$ trained model. This model generally performs best against the union attacks and its performance increases with β, suggesting that the max strategy may be dominated by ℓ_p allowing them to drive up the union's performance.

4.3 Computational complexity

We note that our $\text{TRADES}_{\text{All}}$ model achieves the best overall accuracy while being more computationally efficient than TRADES models trained on RT, $\ell_\infty \circ$ RT, or $\ell_\infty \cup$ RT settings (see CIFAR-10 table. 3). This is because the All strategy only requires computing ℓ_∞ and RT perturbations for $\frac{2}{3}$ of the images every epoch.

5 Discussion

5.1 Takeaways

Our empirical study has three main takeaways: composite attacks are stronger that ℓ_∞, RT, or $\ell_\infty \cup$ RT, complex alternating schemes may be needed to train defenses robust $\ell_\infty \circ$ RT, and robustness tradeoffs exist between specialized and general models. As the results show, all defense methods on both datasets show significant reductions in robust accuracy when defending against composition attacks, while these images appear no different than their RT counterparts (see fig.2). This demonstrates that obtaining truly robust models may be even more difficult than was previously thought. While the problem is certainly very difficult, alternating training schemes seem to help bridge the gap between performance against $\ell_\infty \circ$ RT adversaries and ℓ_∞ adversarial. Our All strategy garners at least some robustness to new threat models, while sacrificing relatively little in terms of robust or natural accuracy, compared to the large number of additional ℓ_∞ hypercubes the models must classify correctly. This gives us hope that the simple All strategy can be improved upon in future work.

5.2 Limitations

While our results provide strong insight into robustness to composite perturbations there are some limitations.

- our proposed solution to the inner maximization problem for RT is limited by its computational inefficiency.
- While being the best of our methods overall, our $\text{TRADES}_{\text{All}}$ model suffers decreased robustness on ℓ_∞ perturbed and natural images.
- Due to computational limitations for the course project, we were unable to scale our evaluation to the ImageNet dataset, or run more extensive Monte Carlo search over different seeds for the experiments reported.

5.3 Future work

As compositional threat models are scantly discussed by previous work, much still remains to be investigated. Some possible directions for future work include an analysis of theoretical tradeoffs between robustness to individual threat models and robustness to their compositions, improving optimization methods for learning robust models, and exploring distribution alignment methods as in [2].

Our analysis shows that for CIFAR-10, the best performing model under the composition attack is $\text{TRADES}_{\text{All}}$. However, as highlighted above, this training strategy reduced both natural accuracy and ℓ_∞ robustness, begging the question: is a tradeoff inherent?. Greater theoretical analysis of this tradeoff is, therefore, needed if we are to understand compositional attacks more deeply. We hypothesize that such analysis may be be realizable by taking a signal processing perspective and viewing RT transformations as members of the affine group $\text{Aff}(2, \mathbb{R})$ acting on an input signal, a perspective largely explored by the geometric deep learning community [4] which has allowed previous work to achieve provable stability to deformations [19, 5, 10].

Another direction is to improve optimization of robust models, which may be attainable through adaptive training adjustments, e.g. similar to curriculum learning [20] but tailored to the adversarial setting, or simply by reducing the difficulty of optimization (e.g., by leveraging function matching [3]). Intuitively, during training a robust model may at any time have learned a function which is more robust to a certain perturbation type than another; it would, therefore, make sense to monitor this performance during training at minimal extra cost and adjust the learner's curriculum accordingly. Alternatively, model distillation is typically leveraged for model compression, but recent results suggest that training a smaller model to match a larger model's logits when fed the same input, a technique coined function matching, allows for theoretically 'infinite' training examples. The large amount of training data could be used to convert a naturally trained model into a distilled model which also minimizes a regularization term, e.g. from equation (5).

Lastly, distribution alignment may also be a promising approach as these different threat models can simply be viewed as different distributions that a robust models must learn. Exploring techniques form the domain adaptation literature similar to [2] but in a more general attack setting could deepen our understanding of the problem and provide a useful defense.

6 Conclusion

Defending against compositional threat models is a difficult but necessary task. While robust models should certainly not need to defend against every possible compositions of adversaries, for instance $\ell_\infty \circ \ell_1$ or $\ell_\infty \circ \ell_2$ make little sense, compositions which leave images imperceptibly altered (fig. 2) but allow for stronger attacks must be considered. Our contributions in this work take a small step towards this goal by highlighting the shortcomings of the ℓ_p threat model, proposing new defense and attack methods for our new but well-motivated threat model, and empirically benchmarking their performance. Our experiments show that alternating training schemes are necessary for striking a balance between different threat models, however, even our best performing method, $\text{TRADES}_{\text{All}}$, does not match the robust performance of specialized models in the different settings considered. These results highlight the need for future research under this threat models both in view of theoretical understanding and empirical defense.

References

[1] Mislav Balunovic, Maximilian Baader, Gagandeep Singh, Timon Gehr, and Martin T. Vechev. Certifying geometric robustness of neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 15287–15297, 2019.

[2] Pouya Bashivan, Reza Bayat, Adam Ibrahim, Kartik Ahuja, Mojtaba Faramarzi, Touraj Laleh, Blake A. Richards, and Irina Rish. Adversarial feature desensitization. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 10665–10677, 2021.

[3] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. *CoRR*, abs/2106.05237, 2021.

[4] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velickovic. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *CoRR*, abs/2104.13478, 2021.

[5] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1872–1886, 2013.

[6] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *CoRR*, abs/2205.08534, 2022.

[7] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1802–1811. PMLR, 2019.

[8] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1802–1811. PMLR, 2019.

[9] Logan Engstrom, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. *CoRR*, abs/1712.02779, 2017.

[10] Shanel Gauthier, Benjamin Thérien, Laurent Alsène-Racicot, Muawiz Chaudhary, Irina Rish, Eugene Belilovsky, Michael Eickenberg, and Guy Wolf. Parametric scattering networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5749–5758, June 2022.

[11] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[12] Sandesh Kamath, Amit Deshpande, Subrahmanyam Kambhampati Venkata, and Vineeth N. Balasubramanian. Can we have it all? on the trade-off between spatial and adversarial robustness of neural networks. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27462–27474, 2021.

[13] Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Geometric robustness of deep networks: Analysis and improvement. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4441–4449. Computer Vision Foundation / IEEE Computer Society, 2018.

[14] Linyi Li, Maurice Weber, Xiaojun Xu, Luka Rimanic, Bhavya Kailkhura, Tao Xie, Ce Zhang, and Bo Li. TSS: transformation-specific smoothing for robustness certification. In Yongdae Kim, Jong Kim, Giovanni Vigna, and Elaine Shi, editors, *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, pages 535–557. ACM, 2021.

[15] Ye Liu, Yaya Cheng, Lianli Gao, Xianglong Liu, Qilong Zhang, and Jingkuan Song. Practical evaluation of adversarial robustness via adaptive auto attack. *CoRR*, abs/2203.05154, 2022.

[16] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[17] Pratyush Maini, Xinyun Chen, Bo Li, and Dawn Song. Perturbation type categorization for multiple adversarial perturbation robustness. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

[18] Pratyush Maini, Eric Wong, and J. Zico Kolter. Adversarial robustness against the union of multiple perturbation models. In *Proceedings of the 37th International Conference on Machine Learning, ICML*

2020, 13-18 July 2020, Virtual Event, volume 119 of *Proceedings of Machine Learning Research*, pages 6640–6650. PMLR, 2020.

[19] Stéphane Mallat. Group invariant scattering. *CoRR*, abs/1101.2286, 2011.

[20] Luke O. Rowe and George Tzanetakis. Curriculum learning for imbalanced classification in large vocabulary automatic chord recognition. In Jin Ha Lee, Alexander Lerch, Zhiyao Duan, Juhan Nam, Preeti Rao, Peter van Kranenburg, and Ajay Srinivasamurthy, editors, *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, pages 586–593, 2021.

[21] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[22] Florian Tramèr and Dan Boneh. Adversarial training and robustness for multiple perturbations. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5858–5868, 2019.

[23] Yun-Yun Tsai, Lei Hsiung, Pin-Yu Chen, and Tsung-Yi Ho. Towards compositional adversarial robustness: Generalizing adversarial training to composite semantic perturbations. *CoRR*, abs/2202.04235, 2022.

[24] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[25] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *CoRR*, abs/2205.01917, 2022.

[26] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *CoRR*, abs/2203.03605, 2022.

[27] 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 Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR, 2019.