Beyond L_p Norms: Delving Deeper into Robustness to Physical Image Transformations

Vikash Sehwag
Princeton University
name of organization (of Aff.)
Princeton, NJ, USA
vvikash@princeton.edu

Jack W. Stokes
Microsoft
Redmond, WA USA
jstokes@microsoft.com

Cha Zhang
Microsoft
Redmond, WA USA
chazhang@microsoft.com

***Abstract*—With the increasing adoption of deep learning in computer vision-based applications, it becomes critical to achieve robustness to real-world image transformations, such as geometric, photometric, and weather changes, even in presence of an adversary. However, earlier work has focused on only a few transformations, such as image translation, rotation, or coloring. We close this gap by analyzing and improving robustness against *twenty-four* different physical transformations. First, we demonstrate that adversarial attacks based on each physical transformation significantly reduce the accuracy of deep neural networks. Next, we achieve robustness against these attacks based on adversarial training, where we show that single-step data augmentation significantly improves robustness against these attacks. We also demonstrate the generalization of robustness to these types of attacks, where robustness achieved against one attack also generalizes to some other attack vectors. Finally, we show that using an ensemble-based robust training approach, robustness against multiple attacks can be achieved simultaneously by a single network. In particular, our proposed method improves the aggregate robustness, against twenty-four different attacks, from 21.4% to 50.0% on the ImageNet dataset.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Deep neural networks have enjoyed large success in a wide range of computer vision applications, such as autonomous driving [1], [2], face recognition [3], [4], robotic vision [5], and healthcare [6], [7]. Given the safety-critical nature of most of these applications, it is essential that deep neural networks are robust to common image transformations in the real world, such as geometric transformation, random noise, image blurring, photometric, and weather changes. Evermore, it is also critical to achieve robustness to worst-case transformations, such that safety cannot be compromised even by a strong adversary.

Though recent works have characterized the performance of state-of-the-art deep neural networks under common image transformations [8], [9], only a limited number of algorithms have achieved robustness against common image transformations in the presence of an adversary [10], [11]. However, the set of transformations considered in these works is limited to geometric transformations, such as translation or rotation, and colorization [11]–[13]. Besides, the set of common image corruptions and perturbations, which a real-world adversary can utilize to craft adversarial examples, is very large. Earlier work, which improves a model's robustness against a few transformations, will still be vulnerable to attacks from multiple other transformations.

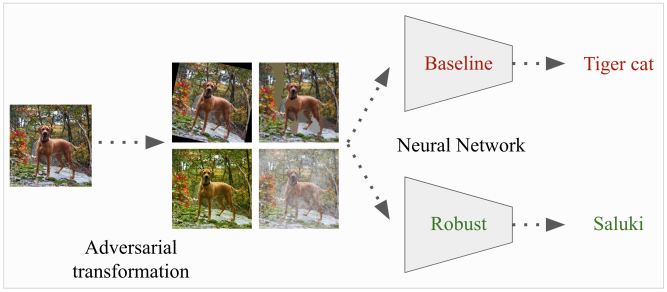


Fig. 1: An illustration of our objective in this paper. We first analyze the performance of deep neural networks under *twenty-four* adversarial transformations. Next we develop *ensemble training* to achieve simultaneous robustness to multiple adversarial transformations.

To close this gap, we aim to both understand and improve the robustness of deep neural networks against a wide range of physical image transformations. We focus on *twenty-four* image transformations from five subgroups including geometric changes, random noise, blurring, photometric modifications, and weather changes. We illustrate our objective in Figure 1 where we first study the robustness of deep neural networks against adversarial examples based on these transformations and later transition to improving the robustness against these attacks.

We first show that deep neural networks remain largely vulnerable to adversarial examples from physical image transformations. For example, coarse random noise-based adversarial examples reduce the test accuracy to as low as 0.1% on the ImageNet dataset (Figure 2). To fix it, we next focus on developing a simple yet effective defense against these attacks. While earlier work on robust training motivates the use of strong adversarial examples [11], i.e, the most adversarial, we find that single-step data-augmentation itself is highly effective in defending against these attacks. It simply requires training with a randomly transformed version of the image in each training step.

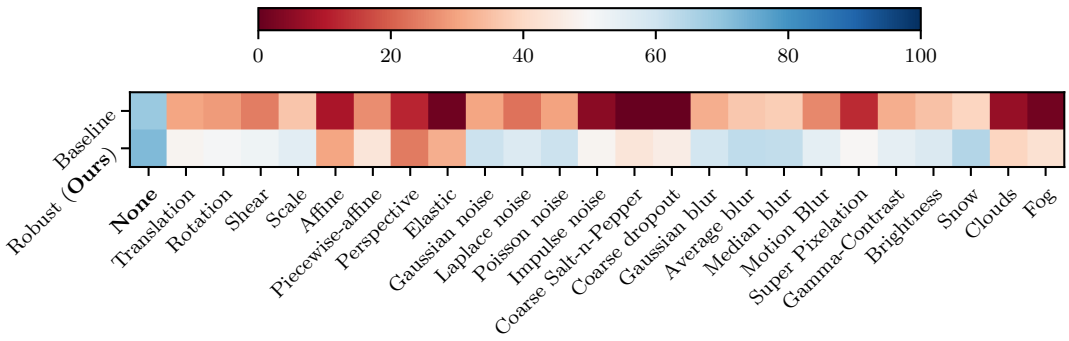


Fig. 2: Test accuracy in the presence of each of the 24 attack vectors for a ResNet-50 network trained on the ImageNet dataset. For non-robustly trained networks (without any data augmentation), i.e., baseline, these attacks reduce the test accuracy from 68.4% to 21.4%. The proposed robust *ensemble training* with all transformations, improves it to 50.0%.

As illustrated in Figure 1, our objective is to train a network which can be simultaneously robust to multiple transformations. When we robustly train a network against an individual transformation, it still remains vulnerable to multiple other transformations, especially ones from different subgroups. We argue that a successful defense will need to accommodate multiple transformations in the robust training objective itself. However, related work in L_p norm-based robust training has shown that multiple adversaries often conflict [14] and can also lead to large degradation of benign accuracy [15].

Our investigation demonstrates that a simple approach of *ensemble training*, where performing robust training against a set of transformations, choosing one randomly for each image in every training step, is highly successful in improving the robustness to multiple transformations, simultaneously. On the ImageNet dataset, it improves the mean robust accuracy across all transformations from 21.4% to 50.0% (Figure 2). We also show that unlike L_p norm-based objectives, multiple objectives in ensemble training do not conflict with each other and do not lead to a degradation of benign accuracy, i.e, accuracy on unmodified test images.

Contributions. We make the following key contributions.

- We conduct a thorough evaluation of adversarial attacks based on twenty-four different physical image transformations, from five different subgroups, on the ImageNet and Reduced-ImageNet dataset.
- We demonstrate that single-step data augmentation can achieve high success in defending against attacks based on physical image transformations. We also demonstrate that there exists some generalization of robustness to unknown transformations within a subgroup or from other subgroups.
- We further show that robustness against multiple attacks can be achieved simultaneously by a single network, without degradation of benign accuracy.

II. Related work

Azulay and Weiss [8] earlier observed that deep neural networks are vulnerable to small geometric changes in the image while Hendrycks and Dietterich [9] extended this observation to multiple other transformations. However, these works only focus on an average case robustness of deep neural networks.

A recent line of research focuses on the robustness against worst-case transformation, i.e., adversarial examples based on physical image transformation. With a motivation to move beyond L_p-norm-based adversarial examples [16], some earlier works have constructed adversarial examples based on translation, rotation, and colorization [10]–[13], [17], [18]. A few other related works add adversarial patches on the image [19]–[21]. We complement that line of work by demonstrating that our twenty-four physical image transformations are highly adversarial thus significantly reducing the accuracy of deep neural networks.

Along with adversarial attacks, several earlier works have also focused on developing robust defenses against translation and rotation based on robust training [11], [22]. A different set of works aims to achieve provable robustness but mostly limited to a few geometric transformations [23], [24]. We show that instead of relying on a computationally expensive robust training framework used in earlier works, a single-step data augmentation in robust training can be highly successful in improving robustness against image transformations.

Another related line of work focuses on robustness to few geometric transformation based on advances in network architectures [25]. We achieve robustness to multiple attacks using robust training.

III. Generating adversarial examples based on physical image transformations

In this section, we first provide a background of adversarial examples and our notation. Next, we present the mechanism used to generate adversarial examples in this work.

Notation. We denote neural network parameters by θ and assume that the training and test inputs are i.i.d. samples from the data distribution, i.e, $(x,\ y) \in \mathbb{D}_{in}$, where x is the input image and y is the respective output class label. We train the neural network $f(\theta, .) : \mathcal{X} \rightarrow \mathcal{Y}$ by minimizing training loss $L(\theta, x, y), \forall (x, y) \in \mathcal{X} \times \mathcal{Y}$ over all training samples. We denote a transformation as $T(\gamma, .) : \mathcal{X} \rightarrow \mathcal{X}$, where γ represents the implicit parameters used by each transformation.

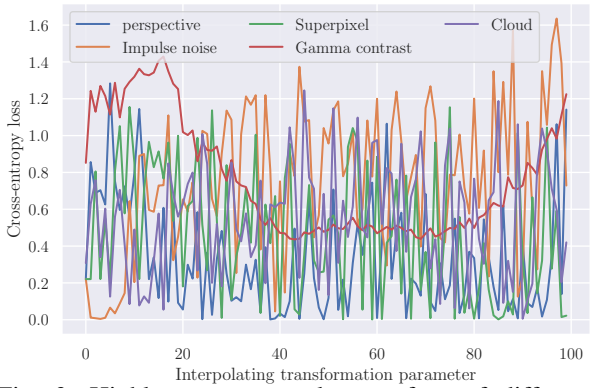


Fig. 3: Highly non-convex loss surface of different transformations, w.r.t. its parameters (γ) for a randomly selected image (ImageNet), which makes it very difficult to generate adversarial examples using gradient-based optimization.

Given an image x and transformation T, we can generate an adversarial example ($\hat{x}$) by solving the following optimization problem.

$$\hat{\gamma} = arg\max_{\gamma} L(\theta, T(\gamma, x), y)$$
$$\gamma \leq \gamma_{max} \quad \hat{x} = T(\hat{\gamma}, x)$$

where $L(.)$ is the required loss function, and γ_{max} is the maximum budget for each transformation, such as limiting rotation to ± 30 degrees. We can use first-order optimization methods, which have been very successful in generating adversarial examples based on pixel-perturbations [26], to solve this problem.

However, we find very limited success when using gradient-based optimization. A key reason behind its failure is the highly non-convex nature of the loss-surface. Engstrom et al. [11] first highlighted this issue when constructing adversarial examples based on image rotation and translation. We find that the loss surface for multiple other transformations, ranging from geometric to weather changes, is also highly non-convex w.r.t. its parameters (Figure 3).

Attacks based on a worst-of-k approach. Most physical image transformations, such as rotation, translation, contrast, and blurring, are parameterized by a few parameters, which makes it possible to search the whole space of parameters. Thus for each image, rather than optimizing based on gradients, we can create a set of transformed images based on each possible parameter setting, and choose the one which is most adversarial. Earlier work has demonstrated that often choosing the most adversarial sample from a set of k randomly sampled parameters (γ) is sufficient to construct successful adversarial examples. This approach is known as the worst-of-k based adversarial attack [11]. We unify adversarial attacks with each transformation by conducting a worst-of-k style attack with it.

We consider a set of 24 physical image transformations to construct adversarial examples (Table I). These transformations capture a wide range of image artifacts, corruptions, and quality degradations observed in the physical world. We categorize these transformations in five subgroups: Geometric, Random Noise, Blurring, Photometric, and Weather transformations. We study these transformations both individually and according to subgroups.

Category	Attack Vectors	Category	Attack Vectors
Geometric Transformations	Translation	Image Blurring	Gaussian Blurring
	Rotation		Average Blurring
	Scale		Median Blurring
	Shear		Motion Blurring
	Affine		Super Pixelation
	Piecewise-Affine	Addition of Random Noise	Gaussian Noise
	Perspective Transform		Laplace Noise
	Elastic Transform		Poisson Noise
Photometric	Gamma and Constrast		Impulse Noise
	Brightness		Coarse Salt-n-Pepper
Weather Simulations	Snow		Coarse Dropout
	Clouds		
	Fog		

TABLE I: Summary of the key image transformations and their subgroups to construct adversarial examples.

IV. Defending against adversarial examples based on physical image transformations

As we demonstrate in Section VI-A, adversarial transformation from each of the transformations in Table I are highly successful in reducing the test accuracy. In this section, we present our approach towards developing a successful defense against physical transformation-based adversarial examples. We focus on a robust training-based defense since it has been highly successful in defending against adversarial examples [11], [15]. It modifies the training objective, where instead of doing empirical risk minimization on the training data, we minimize the training loss on respective adversarial examples. In particular, we solve the following optimization problem:

$$\min_{\theta} \frac{1}{N} \sum_{1}^{N} L(\theta, \hat{x}_t^i, y^i) \tag{1}$$

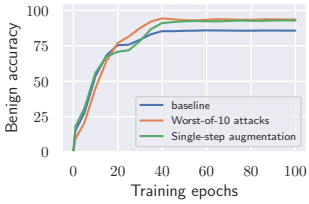
where $\hat{x}$ is the respective adversarial example generated for image x using transformation t. We use the worst-of-k attack to generate each adversarial example.

Primary challenge. Note that a worst-of-k attack requires generating k random transformations of each image. Thus when used in robust training, we will need to evaluate k samples for every image in the batch throughout training. This makes robust training with a worst-of-k attack approximately $k\times$ slower[1] than benign training. Even with a modest value of k, e.g. 10, the computational cost might prohibit its scaling to large-scale datasets, such as ImageNet.

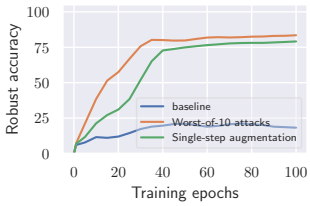
Single-step data augmentation. Our key hypothesis is that given a small parameter space and generalization ability of deep neural networks, robust training with a very weak adversary might suffice to achieve high robustness. In Figure 4, we show that robust training with k=1, for the worst-of-k attack, does indeed achieve high robustness while incurring *negligible*

[1]Not exactly k since forward pass takes less than time than backward pass in deep neural networks.

computational overhead over benign training.[2] Note with k=1, it reduces to using one random transformation of the image, instead of the original image, which we refer to as *single-step data augmentation*.



(a) Benign accuracy



(b) Robust accuracy

Fig. 4: We compare the success of single-step data augmentation with a worst-of-10 attack from [11]. It achieves similar benign and robust accuracy compared to worst-of-10 attacks, while being $6\times$ computationally cheaper than it. We use a worst-of-50 attack to evaluate robust accuracy with ResNet-18 network and the R-ImageNet dataset. Baseline refers to benign training with no data augmentation.

Defending against unknown attacks. A real-world adversary can choose any particular physical transformation to craft an adversarial example. While single-step data augmentation is highly effective in defending against a particular attack, it is unclear how effective it will be in defending against unknown attacks. For example, when a network is robustly trained with rotation-based adversarial examples, will it achieve any robustness against other transformations, from the geometric subgroup or from a different subgroup?

Intra-group generalization of robustness. Given a robust network trained against a particular transformation, it captures the average robust accuracy obtained with attacks within the same transformation subgroup? For robust network $f_{t_a}(\theta)$, which is robust against transformation t_a, we measure intra-group generalization by the following robust accuracy:

$$r_{intra} = \frac{1}{|S_m| - 1} \sum_{1}^{|S_m|} 1(t_a \neq t_i) Acc(f_{t_a}, \hat{X}_{t_i}, Y) \qquad (2)$$

where $t_a, t_i \in S_m$, $Acc(.)$ measures the accuracy of the network on given input samples, $\hat{X}_{t_i}$ refers to adversarial examples generated with transformation t_i and S_m refers to a subgroup of transformations.

Inter-group generalization of robustness. Now we formulate the average robust accuracy obtained with attacks from a different subgroup of transformations. For robust network $f_{t_a}(\theta)$, which is robust against transformation t_a, we measure inter-group generalization with the following robust accuracy:

$$r_{inter} = \frac{1}{|S_k|} \sum_{1}^{|S_k|} Acc(f_{t_a}, \hat{X}_{t_i}, Y); t_a, \in S_m; t_i \in S_k \qquad (3)$$

where $\hat{X}_{t_i}$ refers to adversarial examples generated with transformation t_i from subgroup S_k. We provide experimental results on generalization in Section VI-C.

Defending against multiple attacks simultaneously. Since a real-world adversary can launch adversarial attacks based on any image transformation, an effective defense will require designing a system that is simultaneously robust to multiple attacks. As we will see in Sections VI-B and VI-C, single-step data augmentations suffice to achieve high robustness against any individual transformation and the achieved robustness generalizes to some extent to other transformations. While generalization of robustness between attacks provides some simultaneous robustness to multiple attacks, we need additional techniques to achieve high robust accuracy against multiple, simultaneous attacks.

Ensemble training. A naive solution is to use an ensemble of multiple networks, each being robust to a particular transformation. However, we find that each network does remain vulnerable to transformations unknown to it at traning time. We argue that an effective defense requires *training a single network which is simultaneously robust to multiple transformations.* In particular, we propose to solve the following robust ensemble training objective.

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} L(\theta, \hat{x}_{i,t_i}, y) \; ; \quad t_i \sim \{t_a, t_b, ..., t_n\} \qquad (4)$$

We minimize the training loss over the adversarial examples generated from transformation t_i, which is randomly sampled from the set of possible transformations for each image at each training step. We use the single-step data augmentation to construct x_{t_i}.

In developing an ensemble of robust training objectives, it is critical to consider the following challenges.

- Is there a conflict between the objectives of robust training with multiple attack transformations?
- Will there be high accuracy degradation when embedding robustness to multiple attacks?

As we will see in Section VI-D, we answer both questions negatively. We first show that the proposed robust training is able to achieve simultaneous robustness to multiple attacks. In addition, we also did not observe any degradation in test accuracy. In summary, we find robustness to each transformation, when robustly trained against all, is almost similar to robustness achieved by training against an individual transformation. We hypothesize that the high expressive power of deep neural networks enables this phenomenon.

V. Experimental Setup

We use twenty-four common image transformations, ranging from geometric, random noise, photometric, blurring, and

[2]The data-augmentation cost is not significant due to heavily optimized libraries [27].

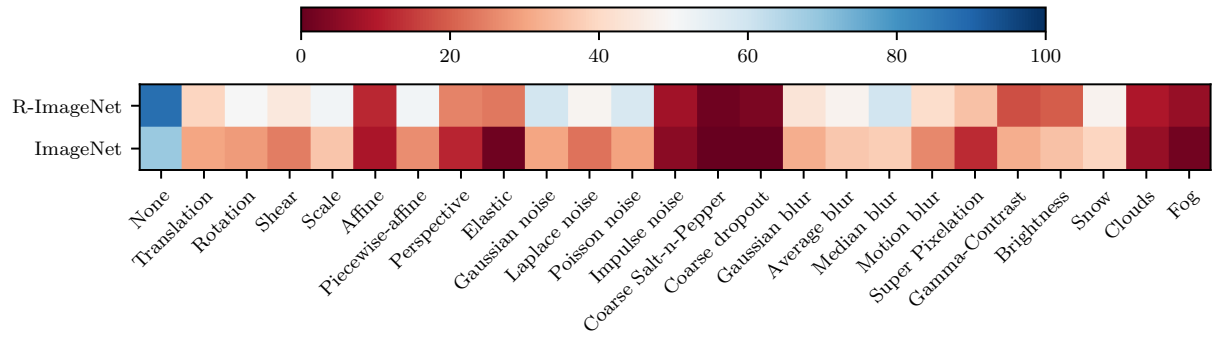


Fig. 5: Robust accuracy against different image transformations, where none refers to benign accuracy. For both dataset, adversarial transformations significantly reduces the accuracy of the network.

weather changes (Table I). To avoid degrading visual quality, we sample parameters of each transformation from a bounded range. We provide the range of these parameters in Appendix 1. We use the well-standardized ImgAug library [27] to implement the transformations. We use the worst-of-50 attack to measure robust accuracy.

We conduct all our experiments on the ImageNet [28] dataset. We deliberately avoid experimenting with very low-resolution datasets, like MNIST and CIFAR-10, where even the smallest physical transformations, such as weather changes or blurring, often heavily distort the image quality. However, to imitate the scale of these smaller datasets, we also experiment with R-ImageNet (Reduced-ImageNet) [29], a 10-class subset of ImageNet. Note that ImageNet has 1000 classes. In comparison to MNIST and CIFAR-10, which have 28×28 and 32×32 size images, respectively, we use 224×224 images for R-ImageNet. We employ widely used Residual networks [30], including ResNet50 for ImageNet and ResNet18 for R-ImageNet, in our experiments. We provide additional experimental details in Appendix 2.

Since our defense is based on data augmentation, we train our baseline networks without any data augmentation. We use the following metrics to evaluate the performance.

Benign accuracy. Measures the percentage of test data points that are classified correctly.

Robust accuracy. Captures the number of correctly classified adversarial examples by a network.

Aggregate robust accuracy. Average of robust accuracy achieved against adversarial examples based on each of the individual transformations.

VI. Experimental Results

We now present results demonstrating the success of adversarial examples based on physical transformations, robustness with single-step based data augmentation and its generalization, and effectiveness of robust ensemble training in achieving robustness to multiple transformations.

A. Success of adversarial attacks based on physical transformations

We present our results in Figure 5 where we report the robust accuracy against each of the twenty-four image transformations for both R-ImageNet and ImageNet dataset. We find that adversarial attacks based on each transformation are highly successful in reducing the accuracy compared to the benign accuracy (i.e., labeled as None).

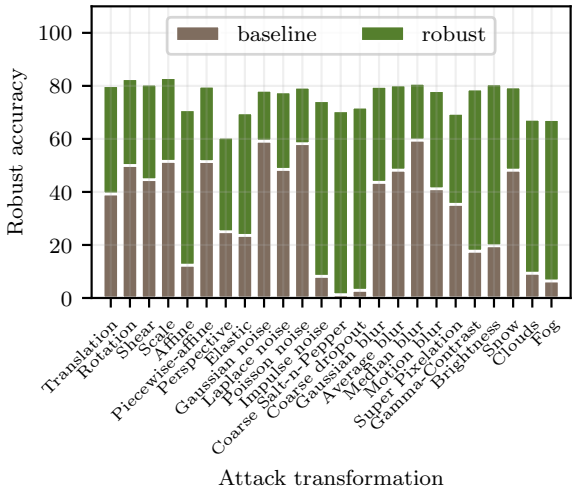
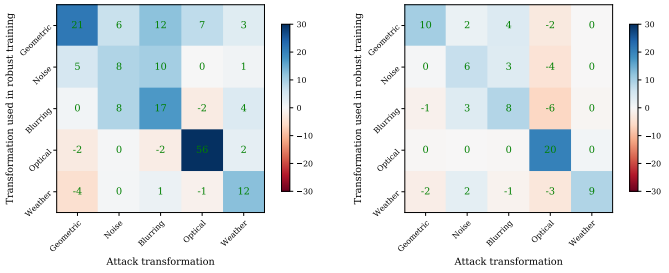


Fig. 6: Robust accuracy against a transform when an individual networks is robustly trained against it on R-Image dataset. We perform robust training with single-step dataset augmentation, which is consistently improving robust accuracy across all transformations. The results for ImageNet dataset are embedded in the Figure 8 itself.

Adversarial attacks are more successful for ImageNet. We observe an aggregated robust accuracy of 33.2% for R-ImageNet and 21.4% for theImageNet dataset. Note that ImageNet includes 1000 classes, in comparison to 10 classes in R-ImageNet, which makes it easier to misclassify images with adversarial perturbation.

Success also depends on the choice of transformation. While most transformations successfully reduce the accuracy of the networks, some are more adversarial than others. For example, perspective transformations, including affine, are generally more adversarial than other geometric transformations. Similarly, coarse noise patterns, such as coarse salt-n-pepper, coarse dropout, and impulse noise, are far more adversarial than other random noise distributions. Finally, we find most weather transformations, such as clouds and fog, are more adversarial than most of the other transformations.



(a) R-ImageNet (b) ImageNet

Fig. 7: Generalization of robust accuracy for both R-ImageNet and ImageNet dataset. Diagonal entries measures intra-group generalization (r_{intra}) and non-diagonal entries measures inter-group generalization (r_{inter}).

Aggregated robust accuracy for subgroups. We observe 37.2%, 29.7%, 46.7%, 18.6%, and 21.3% aggregate robust accuracy for ImageNet and 21.0%, 14.6%, 28.8%, 33.7%, and 15.7% aggregate robust accuracy for R-ImageNet over geometric, random noise, blurring, photometric, and weather subgroups, respectively. Some subgroups, such as weather transformations, are highly adversarial, which necessitates the development of a defense against these attacks.

B. Effectiveness of single-step based data augmentation

Single-step data augmentation with an individual transformation is highly effective in improving a model's robust accuracy when trained for that specific transformation (Figure 6). For example on the R-ImageNet dataset, a ResNet18 network, robust training with rotation-based data augmentation, achieves 82.7% robust accuracy, whereas the baseline model yields only 50% robust accuracy, against rotation-based adversarial examples. Similar improvements are achieved across different transformations, when robust training is performed with them, yielding a mean improvement of 42.6 percentage points. Across the geometric, random noise, blurring, photometric, and weather subgroups, robust training achieves an improvement of 38.7, 45.7, 40.9, 61.0, and 50.1 percentage points in robust accuracy, respectively. Similarly, for the ImageNet dataset, it achieves an improvement of 27.8, 37.3, 26.7, 26.8, and 37.7 percentage points in robust accuracy across the same subgroups, respectively (Figure 8).

C. Generalization of robustness to unseen attacks

While robust training with single-step data augmentation using a particular transformation achieves high robustness against it, it is unclear how well it will generalize, i.e., resist attacks based on other transformations. We provide results on generalization in Figure 7. We study both the inter- and intra-group generalization of adversarial robustness. Against each adversarial transformation, we use the difference of robust accuracy for a robust and the baseline networks, and aggregate it along subgroups based on Equation 2, 3.

Strong intra-group generalization. We find that robustness achieved against a specific transformation also generalizes strongly to others within the same subgroup. For example, a ResNet-18 network which is robustly trained against rotation-based attacks, also improves the robust accuracy by 21 percentage points against other geometric transformations for the R-ImageNet dataset. Similarly, robustness against fog-based transformation also improves robustness to other weather-based transformations by 12 percentage points. We observe a similar trend for the ImageNet dataset (Figure 7b) where robustness to intra-group transformations increases by 6-20 percentage points.

Weak inter-group generalization. We find only a weak generalization of robustness to inter-group transformations. For example, a network robust to rotation-based attacks, only achieve 2-5% improvements in robust accuracy across other transformations from other subgroups. We also observe an inverse generalization where robust training against perturbations from geometric, noise, blurring, and weather subgroups lead to a decrease in the robust accuracy for attacks based on optical changes on the ImageNet dataset.

Dependence on dataset. We find a weaker generalization of robustness for the ImageNet dataset, in comparison to R-ImageNet. It could be attributed to the 1000-class classification for ImageNet, in comparison to 10-class classification for R-ImageNet. In particular, the inter-group generalization degrades most when transitioning from R-ImageNet to ImageNet. For example, an average increase in the inter-group robust accuracy, i.e, mean across all non-diagonal entries in Figure 7, is only -0.25% for ImageNet, whereas it is 2.3% for the R-ImageNet dataset.

D. Simultaneous robustness to multiple attacks

While intra-group generalization of robustness provides some robustness to transformations within the same subgroup, it fails to achieve simultaneous robustness to multiple other attacks. Now we provide results with proposed ensemble training using single-step adversarial examples from all transformations. We present our results in Figure 2, 8.

We demonstrate that robust training with adversarial examples from multiple transformations can enable the network to achieve simultaneous robustness to each of them. While the baseline network achieves only 21.4% aggregate robust accuracy, our robust network achieves 50.0% aggregate robust accuracy, an improvement by more than $2\times$. The corresponding improvement in each subgroup is 21.0, 38.1, 28.2, 22.1, and 32.9 percentage points, respectively. We argue that the very high-expressive power of neural networks enables them to achieve such simultaneous robustness to a range of multiple attacks.

Impact on benign accuracy. Even when robust training with multiple transformations, we find that the benign accuracy of the network does not degrade. In contrast, using single-step data augmentation improves a network's generalization and itself improves the benign accuracy. We find this effect in existent even when doing data augmentation from a set of multiple transformations. We find that the robustly trained networks with all transformations achieve 71.8% benign accuracy while

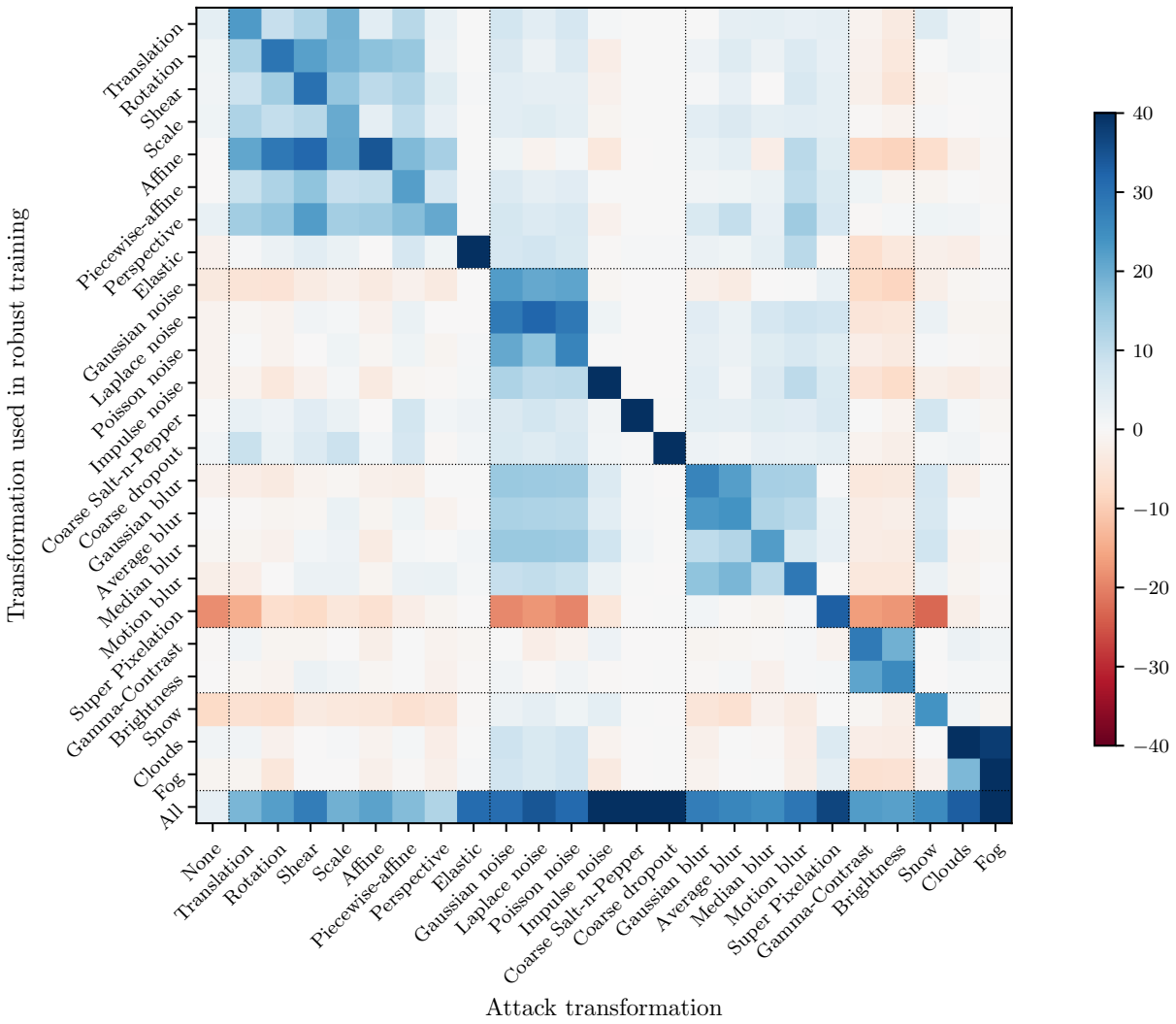


Fig. 8: Detailed results for each of attack-defense pair on ImageNet dataset. For each transformation we robustly train an individual network and measure its accuracy with each attack transformation. *All* refers to robust *ensemble training*.

baseline networks, trained without any data augmentation, only achieve 68.4% benign accuracy.

Most adversarial transformations. Our results suggest the trend that transformations that are most adversarial against a baseline network are harder to defend against. Certain transformations, such as coarse random noise, clouds, fog, affine, perspective, and elastic transformations, remain most adversarial for both baseline and robustly trained networks.

VII. Discussion and Conclusion

We demonstrated that a deep neural network can achieve simultaneous robustness to multiple physical transformation-based attacks. However, do these networks have high enough expressive power to also achieve additional robustness against L_p norm-based adversarial attacks? We answer this question affirmatively. We find that including L_∞-perturbation with existing ensemble training setup leads to only 1%-5% degradation in robustness to physical transformations, while achieving 64% robustness in L_∞-perturbation-based attacks. Ensemble training with only physical transformations leads to 0% robust accuracy against the L_∞ perturbation for 2/255 for ResNet18 network and R-ImageNet dataset.

Conclusion. Deep learning is making rapid progress in many safety-critical areas that rely on images and video such as autonomous driving and healthcare. Robust training has shown to be effective in preventing adversarial learning-based attacks and limited physical attacks for these types of media. We extend earlier work and consider adversarial learning attacks based on twenty-four types of image transformations. We demonstrate that these adversarial attacks are highly effective, even at the scale of the ImageNet dataset. Next, we demonstrate that single-step data augmentation algorithm can provide robustness to individual transformations, with a negligible computation overhead over benign training. We show that the achieved robustness also generalizes to some other attack vectors. Finally, we demonstrate that a single network can be effectively trained to handle multiple, simultaneous attacks. This work is an important step towards protecting a user's

safety in critical applications that rely on image classification.

REFERENCES

[1] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2722–2730.

[2] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[3] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015.

[4] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep face recognition: A survey," in *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*. IEEE, 2018, pp. 471–478.

[5] I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas *et al.*, "Solving rubik's cube with a robot hand," *arXiv preprint arXiv:1910.07113*, 2019.

[6] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[7] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[8] A. Azulay and Y. Weiss, "Why do deep convolutional networks generalize so poorly to small image transformations?" *arXiv preprint arXiv:1805.12177*, 2018.

[9] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019.

[10] C. Kanbak, S.-M. Moosavi-Dezfooli, and P. Frossard, "Geometric robustness of deep networks: analysis and improvement," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4441–4449.

[11] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, "Exploring the landscape of spatial robustness," in *International Conference on Machine Learning*, 2019, pp. 1802–1811.

[12] H. Hosseini and R. Poovendran, "Semantic adversarial examples," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1614–1619.

[13] C. Laidlaw and S. Feizi, "Functional adversarial attacks," in *Advances in neural information processing systems*, 2019, pp. 10 408–10 418.

[14] F. Tramèr and D. Boneh, "Adversarial training and robustness for multiple perturbations," in *Advances in Neural Information Processing Systems*, 2019, pp. 5866–5876.

[15] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[16] M. Sharif, L. Bauer, and M. K. Reiter, "On the suitability of lp-norms for creating and preventing adversarial examples," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1605–1613.

[17] H.-T. D. Liu, M. Tao, C.-L. Li, D. Nowrouzezahrai, and A. Jacobson, "Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer," *arXiv preprint arXiv:1808.02651*, 2018.

[18] J. Chen, D. Wang, and H. Chen, "Explore the transformation space for adversarial images," in *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, 2020, pp. 109–120.

[19] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "A general framework for adversarial examples with objectives," *ACM Transactions on Privacy and Security (TOPS)*, vol. 22, no. 3, pp. 1–30, 2019.

[20] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *arXiv preprint arXiv:1712.09665*, 2017.

[21] C. Xiang, A. N. Bhagoji, V. Sehwag, and P. Mittal, "Patchguard: Provable defense against adversarial patches using masks on small receptive fields," *arXiv preprint arXiv:2005.10884*, 2020.

[22] F. Yang, Z. Wang, and C. Heinze-Deml, "Invariance-inducing regularization using worst-case transformations suffices to boost accuracy and spatial robustness," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 785–14 796.

[23] M. Balunovic, M. Baader, G. Singh, T. Gehr, and M. Vechev, "Certifying geometric robustness of neural networks," in *Advances in Neural Information Processing Systems*, 2019, pp. 15 313–15 323.

[24] M. Fischer, M. Baader, and M. Vechev, "Certification of semantic perturbations via randomized smoothing," *arXiv preprint arXiv:2002.12463*, 2020.

[25] B. Dumont, S. Maggio, and P. Montalvo, "Robustness of rotation-equivariant networks to adversarial perturbations," *arXiv preprint arXiv:1802.06627*, 2018.

[26] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[27] A. B. Jung, K. Wada, J. Crall, S. Tanaka, J. Graving, C. Reinders, S. Yadav, J. Banerjee, G. Vecsei, A. Kraft, Z. Rui, J. Borovec, C. Vallentin, S. Zhydenko, K. Pfeiffer, B. Cook, I. Fernández, F.-M. De Rainville, C.-H. Weng, A. Ayala-Acevedo, R. Meudec, M. Laporte *et al.*, "imgaug," https://github.com/aleju/imgaug, 2020, online; accessed 01-Feb-2020.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

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

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.