Invariance-inducing regularization using worst-case transformations suffices to boost accuracy and spatial robustness

Anonymous

Abstract

This work provides evidence that adding an invariance-inducing regularizer to standard and adversarial training increases predictive accuracy for worst-case spatial transformations (spatial robustness). In fact, with the same computational budget, it achieves a relative reduction of the spatially robust error of 20% for CIFAR10, even surpassing hand-crafted spatial-equivariant networks. For SVHN, we additionally observe that regularized training improves both the standard test and robust accuracy against spatial transformations.

1 Introduction

As deployment of machine learning (ML) systems in the real world has steadily increased over recent years, more and more emphasis is placed on the reliability of the algorithms. It is for example important to understand the properties of commonly used neural networks including invariances to certain types of perturbations from both a security and interpretability point of view.

In this paper we want to explore invariance against spatial transformations like rotations and translations. For deep neural networks, invariance against naturally occurring *spatial transformations* like rotations and translations has been targeted by carefully designing architectures to fulfill these properties [9, 13, 5, 16, 29, 27, 7, 25]. In addition to a handcrafted layer design, the training data is often also *augmented*, with a different random transformation of each image is fed to the optimization algorithm in every epoch [2, 30].

The idea of using data augmentation to enforce invariance properties has also been used in the field of adversarially robust learning [24, 18, 12, 20, 4, 15, 22, 23, 21, 28, 17]. The goal there is to obtain accurate predictions even under worst-case pixel-level perturbations (*attacks*) that are small in some ℓ_p norm and are impercetible to humans. Models such as neural networks have been shown to fail drastically under such attacks despite achieving high natural test accuracy in many applications.

Given the above two lines of work, we want to explore the following questions: (a) how far can we push augmentation based methods for common deep neural networks to be more robust against spatial transformations? (b) how robust are specialized architectures designed to achieve invariance against entire transformation groups? and (c) does robust optimization against spatial transformations necessarily lead to a model with lower accuracy on the unmodified test set (*natural accuracy*)?

In order to explore (a) and (b) we implemented various training methods with an explicit invariance-inducing regularization as described in Section 2 and evaluate spatial robustness for augmentation based methods and specialized neural network architectures on CIFAR10 and SVHN test images. We find that regularized methods can achieve $\sim 20\%$ relative adversarial error reduction compared to previously proposed augmentation-based methods (including adversarial training) without requiring additional computational resources. They empirically even outperform a few traditional spatial-equivariant networks on CIFAR10 and SVHN. Finally, we observe that on SVHN, not only does the robust test accuracy increase with invariance-promoting regularization but it helps to boost standard accuracy as well. We provide theoretical justifications for effectiveness of regularization for both spatial robustness and standard accuracy in Section A.1.

Preprint. Under review.

2 Experimental setup

We now present the details of our methods that we compare in terms of robustness against spatial transformations.

2.1 Regularization based methods

Denoting by $\mathcal{B}(x)$ the perturbation (also called *transformation invariant*) set of images that result from translations of maximally 3 pixels and rotations of at most 30 degrees, we consider SGD applied to the following loss functions

$$\begin{aligned}\min_{f\in\mathcal{F}} \mathcal{L}_{nat}(f;R,\lambda) &:= \min_{f\in\mathcal{F}} \mathbb{E}\, \ell(f(X),Y) + \lambda R(f,X,Y)\\ \min_{f\in\mathcal{F}} \mathcal{L}_{rob}(f;R,\lambda) &:= \min_{f\in\mathcal{F}} \mathbb{E} \sup_{x'\in G^X} \ell(f(x'),Y) + \lambda R(f,X,Y). \end{aligned} \tag{1}$$

where the expectation is with respect to the empirical distribution and we consider the following different choices for R that are encouraging invariance

- $R_{\ell_2}(f,X,Y) = \sup_{x'\in\mathcal{B}(x)} \|f(X)-f(x')\|_2^2$
- $R_{\text{KL}}(f,X,Y) = \sup_{x'\in\mathcal{B}(x)} \mathrm{D}_{\text{KL}}(f(x'),f(X))$ (same as [31] for $\mathcal{L}_{\text{nat}}$)
- $R_{\text{ALP}}(f,X,Y) = \|f(x')-f(X)\|_2^2$ with $x' = \arg\max_{u\in\mathbb{G}^X} \ell(f(u),Y)$ (ALP[10])

We describe the different methods we used to calculate the supremum over the perturbation set in Section B.1 under *defenses* and refer to the maximizers of both regularizers and cross entropy loss as *adversarial examples*.

2.2 Attacks and defenses

The attacks and defenses we choose largely follow the setup in [6]. Recall that the defense refers to the (efficient) procedure at training time which aims to make the resulting model robust to adversarial examples and generally differs from the (extensive) attack mechanism performed at evaluation time to assess the model's robustness.

Grid search attack Following [6], during test time we use grid search with 775 distinct combinations of translation and rotations to find the most adversarial example in the perturbation set for each test image. Although projected gradient descent methods are effective in increasing adversarial test accuracy for ℓ_∞ perturbations, we confirm previous observations [6] that they are worse than grid search in this small dimensional perturbation set.

Defenses For the adversarial example which maximizes either the loss or regularization function, we use the following defense mechanisms:

- Wo-10: the worst of ten randomly drawn transformation
- S-PGD: Projected Gradient Descent with respect to the transformation parameters.
- Random: A randomly drawn transformation

More details about the attack and defense methods can be found in Section B.1 in the Supp. Mat..

For clarity of presentation of our different methods we use a naming convention consisting of the following components: (a) `Reg` : refers to what regularizer was used (ALP, ℓ_2, KLor AT) where ATrefers to standard unregularized adversarial training(b) `batch`: refers to using the gradient of the loss on the adversarial examples (rob), natural examples (nat) or both (mix), and (c) `def`: the mechanism used to find the adversarial example, including random (rnd), worst-of-10 (Wo-10) and spatial PGD (S-PGD) as described in Section B.1. Thus, `Reg` (`batch`, `def`) corresponds to using `Reg` as a regularizing function, the examples defined by `batch` in the gradient of the loss and the defense `def` to find the augmented (adversarial) examples.

2.3 Spatial equivariant networks

We compare the robust prediction accuracies from the regularized networks with three specialized architectures G-ResNet, ETN and STN, designed to be equivariant against spatial transformations and translations. (see Section B.2 in the Supp. Mat. for more details).

Table 1: Mean accuracies of models trained with various forms of regularized adversarial training as well as standard augmentation techniques (top) and spatial equivariant networks (bottom). $\star$ denotes standard augmentation plus random rotations.

	std	std*	AT (rob, Wo-10)	KL (rob, Wo-10)	ℓ_2 (rob, Wo-10)	ALP (rob, Wo-10)	KL-C (mix, S-PGD)	ALP (rob, S-PGD)
SVHN (nat)	95.48	93.97	96.03	96.13	96.53	96.30	96.14	96.11
(grid)	18.85	82.60	90.35	92.71	92.55	92.04	92.42	92.32
CIFAR (nat)	92.11	89.93	91.76	90.41	90.53	90.11	89.98	89.85
(grid)	9.52	58.29	71.17	77.47	77.06	75.9	78.93	77.80

	GRN	GRN*	ETN	ETN*	STN	STN*
SVHN (nat)	96.07	95.05	95.53	95.57	95.61	95.55
(grid)	25.12	84.9	13.15	84.21	36.68	79.28
CIFAR (nat)	93.39	93.08	–	–	–	–
(grid)	16.85	71.64	–	–	–	–

3 Empirical Results

This section is devoted to comparing the natural test accuracy (standard accuracy on the test set) and test grid accuracy (as defined in Section B.1) achieved by standard, regularized (adversarial) training techniques and specialized spatial equivariant architectures described in Section 2.3.

In Table 1 we report results for a subset of the experiments to facilitate comparisons. Results for a more complete set of methods can be found in Supp. Mat., Tables 7–11. For each method, we report averages (standard errors are contained in Tables 7–11) computed over five training runs with identical hyperparameter settings. We compare all methods by computing absolute and relative error reductions (defined as $\frac{\text{absolute error drop}}{\text{prior error}}$). It is insightful to present both numbers since the absolute values vary drastically between datasets.

The experiments are conducted choosing deep neural networks as the function space $\mathcal{F}$ and ℓ is the cross-entropy loss. In the main paper we consider the datasets SVHN [19] and CIFAR-10. For the regularized methods, we train a ResNet-32 [8], implemented in TensorFlow [1]. The precise training procedure is described in Sec. B.3. For training with adversarial examples we augment the data using random lef-tright flips and add random translations of ± 4px for non-adversarially trained models. .

Effectiveness of regularization for adversarial training In Table 1 (top), the three leftmost columns represent unregularized methods which all perform worse in grid accuracy than regularized methods. When considering the three regularizers with the same `batch` and `def` (here chosen to be "rob" and Wo-10) regularized adversarial training improves the grid accuracy from 71.17% to 77.47% on CIFAR-10 and 90.35% to 92.71% on SVHN, corresponding to a relative error reduction of 22% and 24% respectively.

Computational considerations In Figure 1, we plot the grid accuracy vs. the runtime (in hours) for a subset of regularizers and defense mechanisms on CIFAR-10 for clarity of presentation. How much overhead is needed to obtain the reported gains? Comparing AT(rob, Wo-k) and ALP(rob, Wo-k) shows that significant improvements in grid accuracy can be achieved by regularization with only a small computational overhead. What if we make the defense stronger? Improving the adversarial example search in unregularized methods by increasing k to 20 is not effective: AT(rob, Wo-20) achieves 72.31% instead of 71.17% grid accuracy for AT(rob, Wo-10) while requiring $3\times$ more training time. Although regularizing with KL generally achieves the best results, the adversarial example search requires two forward

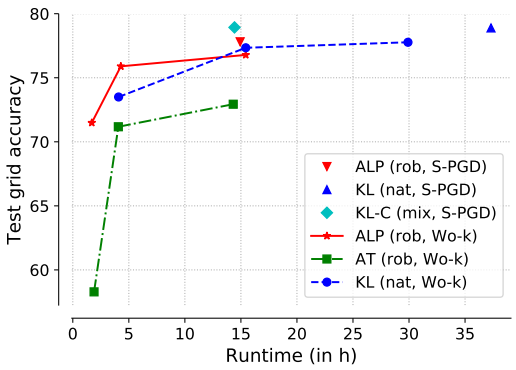

Figure 1: Mean runtime for different methods on CIFAR-10. The connected points correspond to Wo-k defenses with $k \in \{1, 10, 20\}$.

Table 2: Mean accuracies of models trained with various forms of regularized adversarial training. Left: All adversarial examples were found via Wo-10; right: data augmentation setting.

	KL (nat, Wo-10)	ℓ_2 (nat, Wo-10)	ALP (nat, Wo-10)	ℓ_2 (nat, rnd)	KL (nat, rnd)	ℓ_2 (rob, rnd)	KL (rob, rnd)
SVHN (nat)	96.00	96.05	96.39	96.34	96.16	96.09	96.23
(grid)	92.27	92.16	91.98	90.51	90.69	90.48	90.92
CIFAR (nat)	90.83	88.32	88.78	87.80	89.33	88.75	89.47
(grid)	77.34	75.64	75.43	71.60	73.50	71.49	73.22

passes and is hence slow. Hence, we consider
KL-C(mix, S-PGD) which computes the adversarial example with respect to the classification cross entropy loss like ALP. This approach achieves a similar accuracy as KL(nat, S-PGD) while being $2\times$ faster. In fact, this variant achieves the best overall grid accuracy of 78.93% at a similar runtime as AT(rob, Wo-20) with 72.93%, translating into a relative error reduction of 22%.

Comparison with spatial equivariant networks Although the rotation-augmented G-ResNet44 obtains higher grid (84.9%, 71.64%) and natural accuracies ($95\%, 93.08\%$) than the rotation-augmented Resnet-32 on both SVHN ($82.60\%/93.97\%$) and CIFAR10 ($58.29\%/89.93\%$), regularizing augmentation (i.e. regularizers with "rnd", see Table 2 (right)) using both the ℓ_2 distance and the KL divergence matches the G-ResNet44 on CIFAR10 ($71.60\%, 73.50\%$ respectively) and surpasses it on SVHN on grid ($90.51\%, 90.69\%$) and natural accuracies by a relative grid error reduction of $\sim 37\%$. The same phenomenon is observed for the augmented ETN and STN on SVHN. In conclusion, invariance-oblivious regularized augmentation based methods match or outperform state-of-the-art end-to-end networks handcrafted to be equivariant to spatial transformations. relative error

Trade-off natural vs. adversarial accuracy SVHN is one of the main datasets (without artificial augmentation like in rot-MNIST [14]) where spatial equivariant networks have reported improvements on natural accuracy. This is due to the inherent orientation variance in the data. In our mathematical framework, this corresponds to the assumption of the marginal conditional distribution of an transformations of an image having support on the transformation-invariant set. Furthermore, as all numbers in SVHN have the same label irrespective of small rotations of at most 30 degrees, the second assumption in Theorem 2 is also fulfilled. Table 1 confirms the statement in the Theorem that improving robust accuracy may not hurt natural accuracy or even improve it. A similar trend can be seen for CIFAR10 for which the methods with highest grid accuracies such as KL(rob, Wo-10) also have close to the highest natural accuracy.

Comparing the different regularizers and $\mathcal{L}_{\text{nat}}(f; R)$ vs. $\mathcal{L}_{\text{rob}}(f; R)$ We are interested in disentangling the underlying reasons behind differences between the grid accuracy performances of different combinations of regularization methods. Tables 1 and 2 shed some light on this quest. In particular we can observe that, fixing any regularizer, the robust regularized loss `Reg` (rob, `def`) (i.e., $\mathcal{L}_{\text{rob}}(f; R)$) does better (or not statistically significantly worse) than `Reg` (nat, `def`) (i.e., $\mathcal{L}_{\text{nat}}(f; R)$). Second, the KL regularizer is better than ℓ_2 for a large number of settings.

Choice of λ The different regularization methods peak at different λ. However, they outperform unregularized methods in a large range of λ values, suggesting that well-performing values of λ are not difficult to find in practice. These can be seen in Figures 4 and 3 in the Supp. Mat.

Many more insightful experiments can be found in Section B.6.

4 Conclusion

In this work, we have explored how regularized augmented and adversarial training compares against specialized spatial equivariant networks in terms of robustness against spatial transformations consisting of translations and rotations. Strikingly, even though the method of augmentation can be applied to encourage *any* desired invariance, the regularized methods adapt well and perform similarly or better than specialized networks. We showed that for transformation invariances and under certain practical assumptions on the distribution, there is no trade-off between natural and adversarial accuracy which stands in contrast to the debate around ℓ_p-perturbation sets. In summary, it is advantageous to replace unregularized with regularized augmented/adversarial training in a large variety of setings.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Henry S Baird. Document image defect models. In *Structured Document Image Analysis*, pages 546–556. Springer, 1992.

[3] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

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

[5] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pages 2990–2999, 2016.

[6] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning*, 2019.

[7] Carlos Esteves, Christine Allen-Blanchette, Xiaowei Zhou, and Kostas Daniilidis. Polar transformer networks. In *International Conference on Learning Representations*, 2018.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial Transformer Networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[10] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

[11] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 4, University of Toronto, 2009.

[12] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[13] Dmitry Laptev, Nikolay Savinov, Joachim M Buhmann, and Marc Pollefeys. Ti-pooling: transformation-invariant pooling for feature learning in convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 289–297, 2016.

[14] Hugo Larochelle, Dumitru Erhan, Aaron Courville, James Bergstra, and Yoshua Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th International Conference on Machine Learning*, pages 473–480. ACM, 2007.

[15] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[16] Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *ICCV*, pages 5058–5067, 2017.

[17] Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning*, pages 3575–3583, 2018.

[18] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

[19] Y. Netzer, , Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, page 5, 2011.

[20] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 506–519. ACM, 2017.

[21] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, 2018.

[22] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, 2018.

[23] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

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

[25] Kai Sheng Tai, Peter Bailis, and Gregory Valiant. Equivariant Transformer Networks. *arXiv preprint arXiv:1901.11399*, 2019.

[26] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

[27] Maurice Weiler, Fred A Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant cnns. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[28] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5283–5292, 2018.

[29] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5028–5037, 2017.

[30] Larry S. Yaeger, Richard F. Lyon, and Brandyn J. Webb. Effective training of a neural network character classifier for word recognition. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 807–816. MIT Press, 1997.

[31] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. *International Conference of Machine Learning*, 2019.

A Supplementary material

A.1 Theoretical results

In this section, we first introduce our notion of transformation-invariant sets to formalize robustness against a small range of translations and rotations. We then prove that on a population level, constraining or regularizing for transformation invariance yields models that minimize the robust loss. Moreover, under reasonable assumptions on the true distribution, we show the set of robust minimizers not only minimizes the natural loss, but under mild conditions on the distribution over the transformations, is even be equivalent to the set of natural minimizers.

Although the framework can be applied to general problems and transformation groups, we consider the classical image classification case for concreteness. In the following, $X \in \mathbb{R}^d$ are the observed images and $Y \in \mathbb{R}^p$ the one-hot vector for multiclass labels and both are random variables from a joint distribution $\mathbb{P}$. The function $f : \mathbb{R}^d \to \mathbb{R}^p$ in function space $\mathcal{F}$ (e.g. deep neural network in experiments) maps the input image to a logit $x \to f(x)$ that is then used for prediction via a softmax layer.

A.2 Transformation-invariant sets

Invariance against spatial transformations is often thought of in terms of group equivariance of the representation and prediction. Instead of invariance with respect to all spatial transformations in a group, we impose a weaker requirement, that is invariance against transformation-invariant sets.

For any transformation group such as the one consisting of rotations and translations, any compact subset $U \subset \mathcal{X}$ of a group orbit that has $\mathbb{P}(\{X \in U\}) > 0$ is a *transformation-invariant set* (for more details see Section A.6 in the Supp. Mat.). We denote by $\widetilde{\mathcal{G}}$ the set of all transformation-invariant sets for a given transformation group. As a consequence, each $x \in \mathcal{X}$ is uniquely mapped to an element in $\widetilde{\mathcal{G}}$ which we denote by $G^x \in \widetilde{\mathcal{G}}$. We say that a function f is *(transformation-)invariant* if $f(x) = f(x')$ for all $x, x' \in U$ for all $U \in \widetilde{\mathcal{G}}$. Equivalently, the set of all invariant functions can be described by

$$\mathcal{V}(R) := \{f \in \mathcal{F} : R(f, x) = 0 \quad \forall x \in \widetilde{\mathcal{X}}\}$$

where $\mathcal{F}$ is the function space of the models that we are using to fit the data and $R(f, x) = \sup_{x' \in G^x} h(f(x), f'(x))$ for any semimetric[1] h on $\mathbb{R}^p$. For all choices of semimetrics the spaces $\mathcal{V}(R)$ are identical which we denote by $\mathcal{V}$. Using this notation, fitting a model with high accuracy under "small" transformations can be mathematically captured by the robust optimization formulation [3] of minimizing the *robust loss*

$$\mathcal{L}_{\text{rob}}(f) := \min_{f \in \mathcal{F}} \mathbb{E}_{X,Y} \sup_{x' \in G^X} \ell(f(x'), Y). \tag{2}$$

We call the solution of this problem *robust* or the robust minimizer. While adversarial training aims to optimize the empirical version of Eq. (2) directly, the converged predictor might be far from the global minimum of the population optimum in particular in the case of nonconvex landscapes for neural networks. This fact motivates us to explore different formulations in the next section that theoretically result in the same set of minimizers in the population case but exhibit improved robust test accuracy in practice.

A.3 Regularization for adversarial invariance

We consider the following alternative constrained optimization problems of the form

$$\min_{f \in \mathcal{F}} \mathbb{E}\, \ell(f(X), Y) \text{ s.t. } f \in \mathcal{V}(R) \tag{O1}$$

$$\min_{f \in \mathcal{F}} \mathbb{E} \sup_{x' \in G^X} \ell(f(X), Y) \text{ s.t. } f \in \mathcal{V}(R). \tag{O2}$$

They are equivalent to (2) if the set of invariant functions is a subset of the function space $\mathcal{F}$, made rigorous in the following theorem.

[1] a weaker notion of a semimetric satisfies almost all conditions for a metric without having satisfying the triangle inequality

Theorem 1. *If $\mathcal{V} \subset \mathcal{F}$, all minimizers of the adversarial loss* (2) *are in $\mathcal{V}$. If furthermore $\mathcal{V}(R) \subseteq \mathcal{V}$, any solution of the optimization programs* (O1), (O2) *minimizes the adversarial loss.*

The proof of the theorem can be found in the Supp. Mat. in Section A.7. Since projection onto the constrained set is not computationally feasible for neural network function spaces, a common workaround is to relax the constraints by only requiring $f \in \{f : R(f, x) \leq \epsilon \quad \forall x \in \widetilde{\mathcal{X}}\}$ and then use penalization on the natural loss by Lagrangian duality, and rewrite (O1), (O2) as

$$\begin{aligned} \min_{f \in \mathcal{F}} \mathcal{L}_{nat}(f; R, \lambda) &:= \min_{f \in \mathcal{F}} \mathbb{E}\, \ell(f(X), Y) + \lambda R(f, X, Y) \\ \min_{f \in \mathcal{F}} \mathcal{L}_{rob}(f; R, \lambda) &:= \min_{f \in \mathcal{F}} \mathbb{E} \sup_{x' \in G^X} \ell(f(x'), Y) + \lambda R(f, X, Y). \end{aligned} \tag{3}$$

In Section A.5 we discuss how ordinary adversarial training and modified variants that have been proposed thereafter can be viewed as special cases of one of the above formulations (see Section A.5).

A.4 Trade-off between natural and robust accuracy

Even though low robust accuracy (2) might be the number one priority in some applications, one might wonder whether the robust minimizer exhibits higher natural accuracy defined as $\mathcal{L}_{\text{nat}}(f) := \mathbb{E}_{X,Y} \ell(f(X), Y)$. In this section we tackle the final question and identify the conditions in the transformation-invariant setup under which requiring robustness does not hurt or even helps to achieve higher natural accuracy.

The fact that adversarial examples have attracted is partially due to the fact that the predicted label changes for the classifier in a perturbation set in which all images appear the same to the human eye. Mathematically, the latter can be translated as a property of the true label distribution in the setting of invariant set, namely that the conditional distribution Y given x is constant for all x belonging to the same subset $U \in \widetilde{\mathcal{G}}$, or in other words, Y is conditionally independent of X given G^X, i.e. $Y \perp\!\!\!\perp X | G^X$. Under this assumption one can show that for robustness under transformation-invariant sets, there is no trade-off in natural accuracy for the robust minimizer.

Theorem 2 (Trade-off natural vs. robust accuracy)**.** *$Y \perp\!\!\!\perp X | G^X$ holds, the adversarial minimizer also minimizes the natural loss. If moreover, the probability density or mass of X conditioned on every invariant set G^z has support G^z and ℓ is injective, then every minimizer of the natural loss also has be invariant.*

The second statement implies that minimizing the constrained optimization problem (O1) could actually be more efficient in finding the optimal solution to the unconstrained loss minmization, formalized in the theorem below. This could be a possible explanation for why in the empirical section, enforcing the equivariant constraint or adding the penalty causes a boost in both natural and robust accuracy simultaneously.

One might wonder how this result relates to several recent papers such as [31], [26] that presented toy examples for which the ℓ_∞ robust solution must have higher natural loss than the Bayes optimal solution. First of all, the ℓ_∞ sets in their examples are not disjoint and second, the very basic assumption of $Y \perp\!\!\!\perp X | G^z$ that is the reason for surprise in practice is not satisfied.

A.5 Different regularizers and practical implementation

To improve robustness against spatial transformations we consider different choices of $R(f, x)$ in (3) for the regularized version that we then compare in Section 3. Some choices of R correspond to regularized training variants that have been explored in the ℓ_∞ literature. For example, common adversarial training as in [24, 15] corresponds to $\mathcal{L}_{nat}$ with $R(f, X, Y) = \sup_{x' \in G^X} \ell(f(x'), Y) - \ell(f(X), Y)$. In this paper we compare common adversarial training with regularized variants that use the following regularizers

- $R_{\ell_2}(f, X, Y) = \sup_{x' \in G^X} \|f(X) - f(x')\|_2^2$
- $R_{\text{KL}}(f, X, Y) = \sup_{x' \in G^X} \mathrm{D}_{\text{KL}}(f(x'), f(X))$ (TRADES[31] for $\mathcal{L}_{\text{nat}}$) [2]

[2] Notice that TRADES is actually using KL divergence in their code, which is not the same as cross-entropy as claimed in paper.

- $R_{\text{ALP}}(f, X, Y) = \|f(x') - f(X)\|_2^2$ with $x' = \arg\max_{u \in \mathbb{G}^X} \ell(f(u), Y)$ (ALP[10])

where D_{KL} is the KL divergence on the softmax of the vectors f. Standard data augmentation can also be used in conjunction with a regularizer which in the formulation can be modeled by $R(f, X) = \mathbb{E}_{x' \in G^X} h(f, X, X')$ for any semimetric h. Since all of the above regularizers involves a maximization over the invariance set, which is also what is used in practice in the finite sample case, we refer to the maximizer as an *adversarial example*. Note that for the regularizers $R_{\ell_2}, R_{\text{KL}}$ the assumption in Theorem 1 is satisfied and hence the solution of the corresponding constrained problem (O1) is a minimizer of the adversarial loss. In contrast, for ALP, $\mathcal{V}(R) \not\subset \mathcal{V}$ and for data augmentation, $\mathcal{V}(R) \subset \mathcal{V}$ only in the population case.

Adversarial example for spatial transformation sets Since G^X is not a closed group and we do not know whether the drawn $X \in G^z$ for some z lies at the boundary of the group or inside, we cannot solve the maximization constrained to G^X. However, for an appropriate choice of perturbation set (see Section A.6 in the Supp. Mat.), we can instead minimize an upper bound of (2) which reads

$$\min_{f \in \mathcal{F}} \mathbb{E} \sup_{\Delta \in \mathcal{S}} \ell(f(\mathcal{T}(X, \Delta)), Y) \geq \min_{f \in \mathcal{F}} \mathbb{E} \sup_{x' \in G^X} \ell(f(x'), Y). \tag{4}$$

The left hand side is hence what we aim to solve in practice where the expectation is over the empirical joint distribution of X, Y. The relaxation of G^X to the a range of transformations of X that is $\{\mathcal{T}(X, \Delta, :)\Delta \in \mathcal{S}\}$ is also used for the maximization within the regularizers.

A.6 Rigorous definition of invariant sets and choice of $\mathcal{S}$

Let $\mathcal{X}$ be the space of all possible images. We define invariant sets as subsets of transformation groups consisting of all rotations and both horizontal and vertical translations. First we can either view images $I \in \mathcal{X}$ as high dimensional vectors or via a bijective mapping $\Phi : \mathcal{X} \to \mathcal{I}$, as functions $I \in \mathcal{I}$ with $I : \mathbb{Z}^2 \to [0, 255]^3$ mapping from a coordinate to a three dimension vector corresponding to its RGB values. Let us denote by $\mathbb{G}$ the (closed) transformation group acting on $\mathbb{Z}^2$ that consists of elements g_Δ uniquely parameterized by $\Delta \in \mathbb{R}^3$ and can be represented by a coordinate transform matrix, see e.g. [5]. The first two dimensions represent the values for the translations and the third represents translation. We define the transformed image (as an element in $\mathcal{X}$) by group element g_Δ as

$$\mathcal{T}(I, \Delta) := \Phi^{-1}(I(g_\Delta(v_1, v_2)))$$

where v_1, v_2 are the two dimensions of the image. We define as $\mathbb{G}^z$ the $\mathbb{G}$-orbit of z, that is

$$\mathbb{G}^z := \{\mathcal{T}(z, \Delta) : \mathcal{T}(\cdot, \Delta) \in \mathbb{G}\},$$

that is all images that can be obtained from z by a transformation $g_\Delta \in \mathbb{G}$. Hence for all $x, x' \in \mathbb{G}^z$ we have $\mathbb{G}^x = \mathbb{G}^{x'}$ by definition.

On the other hand, the $\mathbb{G}$-orbits partition $\mathcal{X}$, i.e. for $x' \notin \mathbb{G}^x$ we have $\mathbb{G}^x \cap \mathbb{G}^{x'} = \emptyset$ for any pair $x, x' \in \mathcal{X}$. Let's the set of all $\mathbb{G}$-orbit as $\mathcal{G}$. Consequently, for every x there exists a unique element in $U \in \mathcal{G}$ such that $x \in U$. Denoting this surjective mapping by $\Psi : \mathcal{X} \to \mathcal{G}$, we abuse notation a little bit and define $\mathbb{G}^X = \Psi(X)$ as the orbit for which $X \in \mathbb{G}^z$.

Subsets of $\mathbb{G}$-orbits For our problem setting, invariance against transformations that form a closed group constitute too large a set. First, they rarely naturally occur in nature because of physical laws and common human perspectives (an upside down tower for example) and secondly, in image classification robustness is usually only required against adversarial attacks which would not fool a human, i.e. lead to him/her mislabel the image. If the transformation set is too large, this requirement is not fulfilled anymore.

Hence, instead of group invariance on transformation groups with $\mathbb{G}$-orbits, we consider "invariance" on compact subsets of orbits $G^z \subset \mathbb{G}^z$ which we call the *invariant sets*. Given a subset G^z, by definition of the orbits and assumption of compactness of G^z we can define a set of "maximal" transformations $\mathcal{S}_z = (-\Delta_z^\star, \Delta_z^\star)$ using the diameter

$$\Delta_z^\star = \arg\max_{x', x \in G^z} \|\Delta\|_2 \text{ s.t. } x' = \mathcal{T}(x, \Delta). \tag{5}$$

For the purpose of simplicity, we assume that $\Delta^\star$ is constant for all orbits and hence omit the z which yields $\mathcal{S}$ used in the main text.

Denoting the set of subsets as $\widetilde{\mathcal{G}}$, we still have a surjective mapping from $\widetilde{\Psi} : \mathcal{X} \to \widetilde{\mathcal{G}}$. And define $G^x = \widetilde{\Psi}(x)$. In the example of images, an orbit corresponds to all images that can be obtained by transforming one image by translation and rotations. A subset corresponding to an orbit then corresponds to a compact range of orientations and translations for which we care about correct predictions.

Note that the conventional definitions for invariance or equivariance in the group-sense as in [5] do not carry over to subsets of orbits. Our definition of an invariant function $f : \mathcal{X} \to \mathbb{R}^d$ is instead

$$f(\mathcal{T}(x, \mathcal{S})) = f(x') \qquad \forall x' \in \widetilde{\Psi}(x)$$

for all $x \in \mathcal{X}$.

A.7 Proof of Theorem 1

Please refer to Section A.6 for the necessary notation for this section. Furthermore, define $\mathcal{L}_{\text{adv}}(f) = \mathcal{L}_{rob}(f; 0, 0)$, $\mathcal{L}_{\text{nat}}(f) = \mathcal{L}_{\text{nat}}(f; 0, 0)$.

We prove the theorem by contradiction. Let us assume that $f^{\text{rob}} \notin \mathcal{V}$ for exactly one element $G^z \in \widetilde{\mathcal{G}}$ and the marginal distribution over $\widetilde{\mathcal{G}}$ that can be defined as $P(\{G^X = U\}) = P(\{X \in U\})$ for any $U \in \widetilde{\mathcal{G}}$, is discrete(for simplicity of presentation) and G^z has non-zero probability.

Let's assume that there is at least one invariant set G^z, on which f^{rob} is not constant and collect all different values in the set $A = \{f^{\text{rob}}(x) : x \in G^z\}$ (with cardinality strictly bigger than 1 since f not-constant) and denote the distribution over $x \in G^z$ by P_z. Since G^z is assumed to be a subset of a group orbit and $\mathbb{G}$-orbits partition $\mathcal{X}$, there is a mapping Ψ that maps each $x \in \mathcal{X}$ to a unique invariant set (see Section A.6). This property allows us to lower bound of the robust loss as follows for any $z \in \mathcal{X}$:

$$\begin{aligned}
&\mathbb{E}_{X,Y} \sup_{x' \in G^X} \ell(f(x'), Y) \\
=&\mathbb{E}\,[\sup_{x' \in G^X} \ell(f(x'), Y) | X \notin G^z] \mathbb{P}(\{X \notin G^z\}) + \mathbb{E}_{Y|z}[\sup_{x' \in G^z} \ell(f(x'), Y) | G^z] \mathbb{P}(\{X \in G^z\}) \\
\geq&\mathbb{E}\,[\sup_{x' \in G^X} \ell(f(x'), Y) | X \notin G^z] \mathbb{P}(\{X \notin G^z\}) + \sup_{a \in A} \int H(x, a) dP_z(x) \qquad (6)
\end{aligned}$$

where the function H is specific to G^z and defined as $H(x, a) = \mathbb{E}_{Y|x} \ell(a, Y)$ and the inequality follows from

$$\mathbb{E}_{X|z} \mathbb{E}_{Y|x}[\sup_{a \in A} \ell(a, Y) | X = x] | G^z = \int \mathbb{E}_{Y|x} \sup_{a \in A} \ell(a, y) dP_z(x) \geq \sup_{a \in A} \int \mathbb{E}_{Y|x} \ell(a, Y) dP_z(x).$$

The right hand side is minimized with respect to the set A by choosing $A = \{a^\star\}$ with $a^\star = \arg\min_a \int H(x, a) dP_z(x)$ because setting $f(x) = a^\star$ for all $x \in G^z$ leads to equality in equation (6). Since $\mathbb{P}(\{X \in G^z\}) > 0$ by assumption, choosing $f^\star(x) = a^\star$ for all $x \in G^z$ implies $\mathcal{L}_{\text{adv}}(f^\star) < \mathcal{L}_{\text{adv}}(f^{\text{rob}})$ which contradicts optimality of f^{rob} and thus proves the first statement of the theorem.

For the second statement let us rewrite

$$\begin{aligned}
\mathcal{L}_{\text{adv}}(f) &= \mathcal{L}_{nat}(f) + [\mathcal{L}_{\text{adv}}(f) - \mathcal{L}_{nat}(f)] \\
&= \mathbb{E}\,\ell(f_\theta(\mathcal{T}(X, \Delta')), Y) + \underbrace{\mathbb{E}\,[\max_{\Delta' \in \mathcal{S}} \ell(f(\mathcal{T}(X, \Delta')), Y) - \ell(f(X), Y)]}_{\tilde{R}(f)}
\end{aligned}$$

By the first statement we know that the set

$$F^{\text{rob}} := \{f \in \mathcal{V} : \mathcal{L}_{\text{adv}}(f) \leq \mathcal{L}_{\text{adv}}(f') \quad \forall f' \in \mathcal{F}\}$$

is non-empty. For all $f \in F^{\text{rob}}$, it holds by definition of $\mathcal{V}$ that $\tilde{R}(f) = 0$.

Since $\mathcal{V}(R) \subset \mathcal{V}$, the minimizers $f^{\min}$ of (O1) satisfy $\mathcal{L}_{\text{nat}}(f^{\min}) \leq \mathcal{L}_{\text{nat}}(f)$ for all $f \in \mathcal{V}$. But since in $\mathcal{V}$ we have $\mathcal{L}_{\text{adv}}(f) = \mathcal{L}_{\text{nat}}(f)$ it directly follows that $f^{\min} \in F^{\text{rob}}$. The same argument goes through for (O2) since for all $f \in \mathcal{V}$, we have $\mathcal{L}_{\text{adv}}(f) = \mathcal{L}_{\text{nat}}(f)$. This concludes the proof of the theorem.

A.8 Proof of Theorem 2

Similar to the proof of Theorem 1, we can const minimizer $f^\star$ of the natural loss that is in $\mathcal{V}$ given the assumption that $Y \perp\!\!\!\perp X | G^z$. Hence since on $\mathcal{V}$ both losses are the same, together with Theorem 1 the robust minimizer also minimizes the unconstrained natural loss.

Assume $f^{\text{nat}} \notin \mathcal{V}$ minimizes $\mathcal{L}_{\text{nat}}(f)$, in particular, it is constant on all invariant sets except G^z for some $z \in \mathcal{X}$. Again by existence of a mapping Ψ and hence since by assumption, the joint conditional distribution $P_z(Y, X) = P_z(Y)P_z(X)$, we can write for any f

$$\begin{aligned}\mathcal{L}_{\text{nat}}(f) &= \mathbb{E}_{X,Y}\ell(f(X), Y) \\ &= \mathbb{E}\left[\ell(f(X), Y)|X \notin G^z\right]\mathbb{P}(\{X \notin G^z\}) + \mathbb{E}\left[\ell(f(X), Y)|G^z\right]\mathbb{P}(\{X \in G^z\}) \\ &= \mathbb{E}\left[\ell(f(X), Y)|X \notin G^z\right]\mathbb{P}(\{X \notin G^z\}) + \mathbb{E}_X\mathbb{E}_Y\ell(f(X), Y)\mathbb{P}(\{X \in G^z\}).\end{aligned}$$

Note that

$$\begin{aligned}\mathbb{E}_X\mathbb{E}_Y\ell(f(X), Y)|G^z &= \int H(x)dP_z(x) \\ &\geq \int \min_{x \in G^z} H(x)dP_z(x) = \mathbb{E}_X\mathbb{E}_Y\ell(f^\star(X), Y)|G^z \qquad (7)\end{aligned}$$

Then, setting $f^\star(x) = \min_{x \in G^z} \mathbb{E}_Y\ell(f^{\text{nat}}(x), Y)|G^z$ for all $x \in G^z$ yields $\mathcal{L}_{\text{nat}}(f^\star) = \mathcal{L}_{\text{nat}}(f^{\text{nat}}) \leq \mathcal{L}_{\text{nat}}(f)$ for all $f \in \mathcal{F}$ by definition of f^{nat}.

If additionally the support of P_z is equal to G^z and ℓ is injective, the inequality (7) becomes a strict inequality for $f^{\text{nat}} \notin \mathcal{F}$ and hence we have $\mathcal{L}_{\text{nat}}(f^\star) < \mathcal{L}_{\text{nat}}(f^{\text{nat}})$ which contradicts the definition of f^{nat} being the minimizer of the natural loss.

B Experimental details

B.1 Attacks and defenses

Grid search attack We consider a default grid of 5 values per translation direction and 31 values for rotation, yielding 775 transformed examples that are evaluated for each X_i. We refer to the accuracy attained under this attack as *grid accuracy*.[3]

Defenses **worst-of-k** In worst-of-k (Wo-k) adversarial training, at iteration t, we sample k different perturbations for each image in the batch. The one resulting in the highest function value is chosen to be the maximizer. Most of our experiments are conducted with $k = 10$ consistent with [6] as a higher k only improved performance minimally.

Spatial PGD. In analogy to common practice for ℓ_p adversarial training as in e.g. [24, 15], the S-PGD mechanism uses projected gradient descent with respect to the translation and rotation parameters with projection on the constrained set $\mathcal{S}$ of transformations. We consider 5 steps, starting from a random initialization, with step sizes of $[0.03, 0.03, 0.3]$ (following [6]) for horizontal-, vertical translation and rotation respectively.

Random. Data augmentation with a distinct random perturbation per image and iteration as ubiquitously used in practice. This can be seen as the most naive "adversarial" example as it corresponds to worst-of-k with $k = 1$.

Note that the empty space that results from translating and rotating an image is filled with black pixels (*constant padding*) if not otherwise noted, and reflections in some cases (see Section B.5). See Fig. 2 for an example.

B.2 Spatial-equivariant networks

In our experiments, we report spatial robustness for the following spatial-equivariant networks: (a) G-ResNet44 (GRN) [5] using p4m convolutional layers (90 degree rotations, translations and mirror

[3] A finer grid search attack with a total of 7500 transformations only lead to minor reductions in accuracy compared to the coarse grid (see Table 6).

reflections) on CIFAR10; (b) Equivariant Transformer Networks (ETN) [25], a generalization of Polar Transformer Networks (PTN) [7], on SVHN; and (c) Spatial Transformer Networks (STN) [9] on SVHN. A more comprehensive discussion of previous literature on equivariant networks against translations and rotations can be found in Section **??**. We choose the architectures listed above based on availability of reproducible code and previously reported state-of-the art accuracies for the respective datasets.

We train GRN, STN and ETN end-to-end using standard augmentation as described in Section **??** (std) and random rotations in addition (std*). In order to compute the grid accuracies, the adversarial attacks using grid search are then computed by evaluating the entire architecture. Out of curiosity we also trained a two-stage STN where we train the localization network separately on a supervised randomly transformed set of the training data to predict the transformation parameters treated as labels. Details about the implementation and results can be found in Section B.5.

We had difficulties to train both ETN and STN to higher than 86% natural accuracy for CIFAR10 even after an extensive learning rate and schedule search so we do not report the numbers here.

B.3 Training procedure

All models are trained using a single-GPU on a node equipped with an NVIDIA GeForce GTX 1080 Ti and two 10-core Xeon E5-2630v4 processors.

We run a standard minibatch SGD update with a momentum term with parameter 0.9 and weight decay parameter 0.0002. We use an initial learning rate of 0.1 which is divided by 10 after half and three-quarters of the training steps. Independent of the defense method, we fix the number of iterations to 80000 for SVHN and CIFAR-10, and to 120000 for CIFAR-100. For comparability across all methods, we fix the number of unique original images in each iteration to 64. For the baselines, we additionally trained with a batch size of 128 and report the higher accuracy.

B.4 Example images

One pair of example images is shown in panels (b) and (c) in Fig. 2, where the original image is depicted along with a transformed version $\mathcal{T}(\cdot, \Delta)$ with $\Delta \in \mathcal{S}$.

(a) (b) (c)

Figure 2: Example images and classifications by the Standard model. (a) An image that is misclassified for all rotations in the considered grid. (b) An image that is correctly classified for most of the rotations in the considered grid. (c) One rotation for which the image shown in (b) is misclassified as "airplane".

B.5 Design choices for two-stage STN

Since STNs are known to be sensitive to hyperparameter settings and thus difficult to train end-to-end [25], we apply the following two-stage procedure to simulate its functionality: (1) we first train a ResNet-32 as a localization regression network (LocNet) to predict the attack perturbation separately by learning from a training set, which contains perturbed images and uses the transformations as the prediction targets; (2) at the same time we train a ResNet-32 classifier with data augmentation, namely random translations and rotations; (3) during the test phase, the output of the LocNet is used by a spatial transformer module that transforms the image before entering the pretrained classifier. We refer to this two-stage STN as STN+.

For the classifiers, we take the two trained models for CIFAR10 and SVHN from our previous experiments for AT(rob, rnd). Since we do not expect the regressors to be perfect in terms of prediction capability, there will still be some transformation left after the regression stage. Thus, the classifiers should have a certain degree of robustness. Their training procedure is described in Section B.3.

The regressor is trained on zero padded – suffix (c) – as well as reflect padded inputs – suffix (r) – for comparison. The former possibly yields an unfair advantage of this approach compared to other methods as the neural network can exploit the edges (induced through zero padding) to learn

the transformation parameters. Therefore, we also consider reflection padding to assess the effect of the different paddings on final performance. Nonetheless, zero padding is consistent with the augmentation setting for the end-to-end trained networks and regularized methods and was also the choice considered by [6]. For completeness we also show results when using reflection padding for training LocNet although it lacks comparability with the other methods since attacks should be reflection-padded as well.

[6] additionally analyze a "black canvas" setting where the images are padded with zeros prior to applying the transformation, ensuring that no information is lost due to cropping. Their experiments show that the reduced accuracy of the models cannot be attributed to this effect. Since both versions yield similar results, we report results on the first version of pad and crop choices, having input images of the same size as the original.

In the spatial transformer module we compares two variants of handling the labels predicted by the regressor. We can either tilt the transformed image with the opposite transformation of the predicted labels, which will, under the assumption that the regressor successfully learnt object orientations, turn back the image but potentially result in extra padding space before we feed the images into the classifier. Alternatively, we can subtract the predicted transformation from the attack transformation, then use the remaining transformation as the new "attack transformation". The latter will result in much smaller padding areas, if the regressor is performing well. From the experimental results we do see a big drop if we naively transform images twice. We denote the former method as "naive" and latter as "trick".

For CIFAR-10, this two-stage classifier achieved relatively high grid accuracies. However, the obtained accuracies are still lower than expected, given that the regressor is allowed to learn rotations with a separately trained regressor on the transformed training set. For SVHN we also see a gain compared to adversarial training without regularizer. However, the performance still lags behind the accuracies obtained by the regularizers. The results are summarized in Table 3.

Table 3: Accuracies of two-stage STN (STN+) under different settings. Details are provided in Section B.5.

Dataset	STN+(c) trick	STN+(r) trick	STN+(c) naive	STN+(r) naive
SVHN (nat)	94.92	95.51	94.92	95.51
(grid)	90.95	90.28	64.91	59.68
CIFAR10 (nat)	91.29	90.99	91.29	90.99
(grid)	83.05	84.31	44.88	42.84

B.6 Additional experimental results

In this section we discuss additional experimental results we collected and analyses we performed.

Weakness of first order attack. Table 5 shows the accuracies of various models trained with S-PGD defenses and evaluated against the S-PGD and the grid search attack on all datasets. We observe that the S-PGD attack constitutes are very weak attack since the associated accuracies are much larger than for the grid search attack. In other words, the S-PGD attack only yields a very loose upper bound on the adversarial accuracy. This stands in stark contrast to ℓ_∞ attacks and has first been noted and discussed in [6]. Interestingly, using the first order method as a *defense* mechanism proves to be very effective in terms of grid accuracy. When used in combination with a regularizer this defense yields the largest overall accuracies as shown and discussed in Section 3. Recall that due to computational reasons grid search cannot be used as a defense mechanism. Therefore, the strongest computationally feasible defense does not use the same mechanism as the strongest attack in our setting.

Table 4: Mean accuracies of different models trained with and without defenses and evaluated on the natural test set (column "Natural") and against the grid search attack (column "Grid") on CIFAR-100. Standard errors are shown in parentheses.

	Training \ Evaluation	Natural	Grid
CIFAR-100	std	70.23% (0.18%)	5.09% (0.25%)
	AT(mix, rnd)	72.38% (0.17%)	23.43% (0.47%)
	AT(rob, rnd) $\equiv$ std*	66.62% (0.37%)	28.53% (0.25%)
	ALP(mix, rnd) ($\lambda = 0.75$)	66.54% (0.26%)	**45.97% (0.22%)**
	AT(mix, Wo-10)	73.03% (0.13%)	35.93% (0.24%)
	AT(rob, Wo-10)	68.79% (0.34%)	38.21% (0.10%)
	ALP(mix, Wo-10) ($\lambda = 0.25$)	68.54% (0.27%)	**49.30% (0.33%)**
	AT(mix, Wo-20)	73.33% (0.22%)	38.06% (0.19%)
	AT(rob, Wo-20)	69.15% (0.49%)	40.09% (0.31%)
	ALP(mix, Wo-20) ($\lambda = 0.25$)	68.04% (0.27%)	**49.98% (0.31%)**
	AT(mix, S-PGD)	71.11% (0.37%)	33.40% (0.21%)
	AT(rob, S-PGD)	68.87% (0.19%)	37.87% (0.12%)
	ALP(mix, S-PGD) ($\lambda = 0.5$)	68.44% (0.39%)	**52.58% (0.20%)**

Table 5: Mean accuracies of different models trained with S-PGD defenses and evaluated on the natural test set (column "Natural"), against the S-PGD attack (column "S-PGD") and against the grid search attack (column "Grid") on the SVHN, CIFAR-10 and CIFAR-100 datasets. While the test accuracy for the S-PGD attack is only slightly lower than the natural accuracy in most cases, the grid accuracy is significantly smaller. In other words, the S-PGD attack only yields a loose upper bound on adversarial accuracy while it is a competitive defense in terms of test grid accuracy when used together with ALP.

	Training \ Evaluation	Natural	S-PGD	Grid
SVHN	AT(mix, S-PGD)	96.27% (0.00%)	95.26% (0.04%)	84.81% (0.01%)
	AT(rob, S-PGD)	96.06% (0.10%)	95.46% (0.10%)	87.29% (0.09%)
	ALP(mix, S-PGD) ($\lambda = 2.5$)	96.30% (0.09%)	95.92% (0.13%)	**92.42% (0.20%)**
CIFAR-10	AT(mix, S-PGD)	92.19% (0.23%)	88.84% (0.27%)	64.26% (0.25%)
	AT(rob, S-PGD)	91.83% (0.19%)	89.87% (0.10%)	69.74% (0.27%)
	ALP(mix, S-PGD) ($\lambda = 3$)	89.70% (0.10%)	88.15% (0.21%)	**77.72% (0.35%)**
CIFAR-100	AT(mix, S-PGD)	71.11% (0.37%)	65.01% (0.32%)	33.40% (0.21%)
	AT(rob, S-PGD)	68.87% (0.19%)	65.56% (0.12%)	37.87% (0.12%)
	ALP(mix, S-PGD) ($\lambda = 0.5$)	68.44% (0.39%)	66.04% (0.40%)	**52.58% (0.20%)**

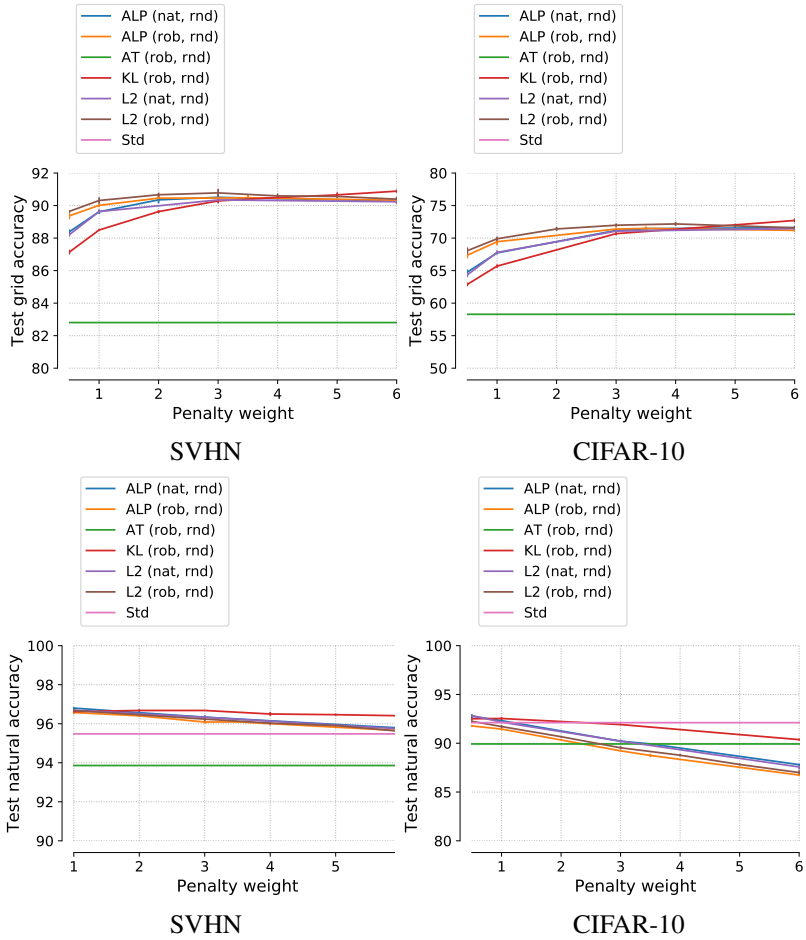

Figure 3: Test grid accuracy (first row) and test natural accuracy (second row) as a function of the regularization parameter λ for the SVHN (first column) and CIFAR-10 (second column) datasets and data augmentation ("rnd"). The test grid accuracy is relatively robust in a large range of λ values while natural test accuracy decreases with larger values of λ.

Table 7: Mean accuracies of models trained *without* regularized adversarial training. Standard errors are shown in parentheses.

	std	std*	AT(rob, Wo-10)
SVHN (nat)	95.48 (0.15)	93.97 (0.09)	96.03 (0.04)
(grid)	18.85 (1.27)	82.60 (0.23)	90.35 (0.27)
CIFAR10 (nat)	92.11 (0.18)	89.93 (0.18)	91.76 (0.23)
(grid)	9.52 (0.66)	58.29 (0.60)	71.17 (0.26)

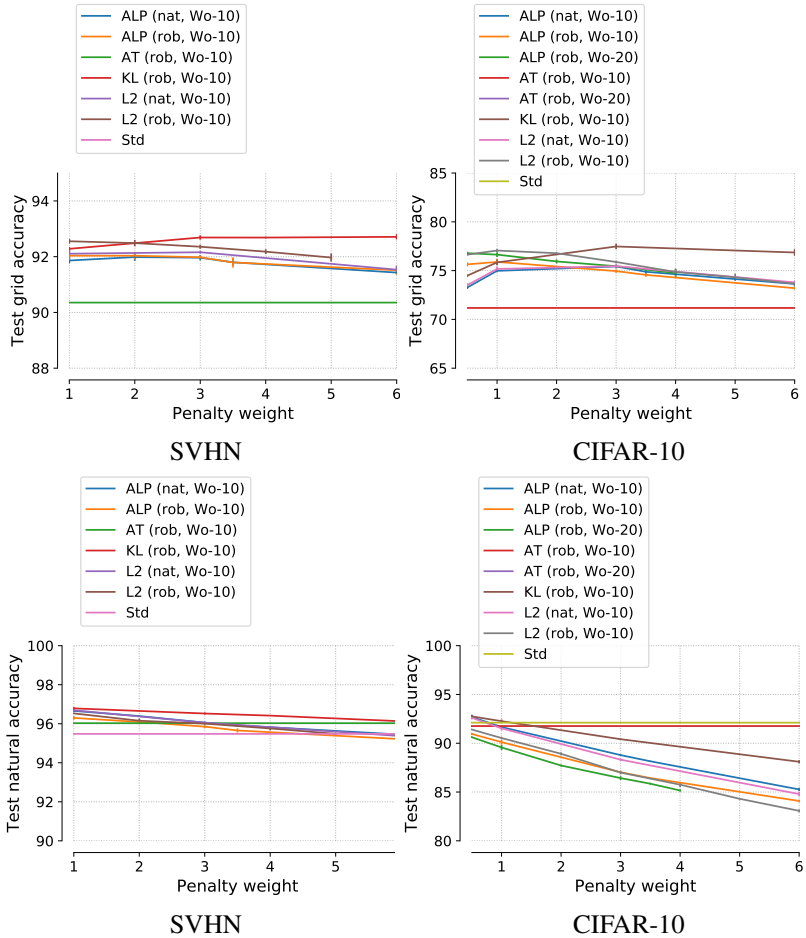

Figure 4: Test grid accuracy (first row) and test natural accuracy (second row) as a function of the regularization parameter λ for the SVHN (first column) and CIFAR-10 (second column) datasets and Wo-k defenses. The test grid accuracy is relatively robust in a large range of λ values while natural test accuracy decreases with larger values of λ.

Table 8: Mean accuracies of models trained with various forms of regularized adversarial training, using the KL regularization function. Standard errors are shown in parentheses.

	KL(nat, rnd)	KL(nat, Wo-10)	KL(rob, Wo-10)	KL-C(mix, S-PGD)	KL(nat, S-PGD)
SVHN (nat)	96.16 (0.10)	96.00 (0.02)	96.13 (0.07)	96.14 (0.04)	96.54 (0.01)
(grid)	90.69 (0.05)	92.27 (0.09)	92.71 (0.09)	92.42 (0.03)	92.62 (0.03)
CIFAR10 (nat)	89.33 (0.16)	90.83 (0.18)	90.41 (0.05)	89.98 (0.21)	89.82 (0.13)
(grid)	73.50 (0.19)	77.34 (0.19)	77.47 (0.28)	78.93 (0.23)	78.89 (0.07)

Table 6: Mean accuracies for different models evaluated against two different grid search attacks. The column "Grid (775)" shows the test grid accuracy using the default grid containing 5 values per translation direction and 31 values for rotation, yielding a total of 775 transformed examples that are evaluated for each X_i. The column "Grid (7500)" shows the test grid accuracy on a much finer grid with 10 values per translation direction and 75 values for rotation, resulting 7500 transformed examples. We observe that the test grid accuracy only decreases slightly for the finer grid and the reduction in accuracy is smaller for ALP than for AT. Due to computational reasons we use the grid containing 775 values for all other experiments.

	Training \ Evaluation	Grid (775)	Grid (7500)
SVHN	AT(mix, Wo-10)	88.83% (0.10%)	88.02% (0.12%)
	AT(rob, Wo-10)	89.75% (0.17%)	89.29% (0.15%)
	ALP(mix, Wo-10) ($\lambda = 1.5$)	92.17% (0.11%)	91.79% (0.12%)
CIFAR-10	AT(mix, Wo-10)	68.14% (0.48%)	65.69% (0.28%)
	AT(rob, Wo-10)	70.35% (0.16%)	68.28%(0.16%)
	ALP(mix, Wo-10) ($\lambda = 1$)	75.89% (0.23%)	74.58% (0.16%)
CIFAR-100	AT(mix, Wo-10)	35.93% (0.24%)	33.62% (0.23%)
	AT(rob, Wo-10)	38.21% (0.10%)	36.04% (0.21%)
	ALP(mix, Wo-10) ($\lambda = 0.25$)	49.30% (0.33%)	47.95% (0.23%)

Table 9: Mean accuracies of models trained with various forms of regularized adversarial training, using the ℓ_2 and ALP regularization functions. Standard errors are shown in parentheses.

	ℓ_2(nat, Wo-10)	ℓ_2(adv, Wo-10)	ALP(mix, Wo-10)	ALP(rob, Wo-10)	ALP(rob, S-PGD)
SVHN (nat)	96.05 (0.04)	96.53 (0.03)	96.41 (0.07)	96.3 (0.09)	96.11 (0.08)
(grid)	92.16 (0.05)	92.55 (0.08)	92.17 (0.11)	92.04 (0.19)	92.32 (0.17)
CIFAR10 (nat)	88.32 (0.13)	90.53 (0.16)	91.13 (0.13)	90.11 (0.25)	89.85 (0.27)
(grid)	75.46 (0.25)	77.06 (0.16)	75.89 (0.23)	75.90 (0.31)	77.80 (0.17)

Table 10: Mean accuracies of models trained with various forms of augmented training, i.e. unregularized and regularized data augmentation. Standard errors are shown in parentheses.

	std*	ℓ_2(nat, rnd)	KL(nat, rnd)	ALP(rob, rnd)	KL(rob, rnd)	ALP(mix, rnd)
SVHN (nat)	93.97 (0.09)	96.34 (0.08)	96.16 (0.10)	96.09 (0.06)	96.23 (0.08)	96.19 (0.07)
(grid)	82.60 (0.23)	90.51 (0.15)	90.69 (0.05)	90.48 (0.16)	90.92 (0.17)	90.48 (0.15)
CIFAR10 (nat)	89.93 (0.18)	87.80 (0.11)	89.34 (0.16)	88.75 (0.18)	89.47 (0.04)	89.43 (0.28)
(grid)	58.29 (0.60)	71.60 (0.27)	73.50 (0.19)	71.49 (0.30)	73.22 (0.14)	71.97 (0.11)

Table 11: Mean accuracies of models trained with various forms of regularized adversarial training. Standard errors are shown in parentheses.

	ALP(nat, Wo-10)	ℓ_2(nat, Wo-10)	KL-C(nat, Wo-10)	KL(nat, Wo-10)
SVHN (nat)	96.39 (0.03)	96.05 (0.04)	96.18 (0.06)	96.00 (0.02)
(grid)	91.98 (0.13)	92.16 (0.05)	91.99 (0.12)	92.27 (0.09)
CIFAR10 (nat)	88.78 (0.11)	88.32 (0.13)	89.61 (0.09)	90.83 (0.18)
(grid)	75.43 (0.13)	75.46 (0.25)	76.15 (0.23)	77.34 (0.19)

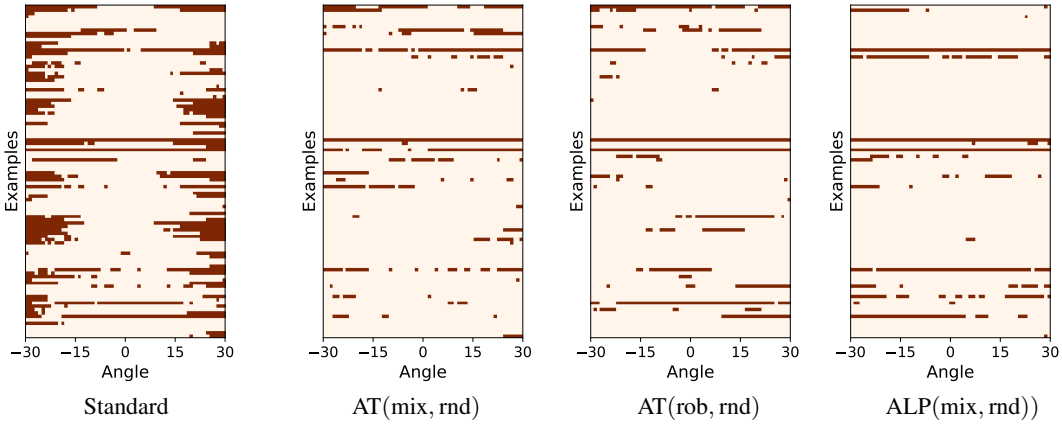

Figure 5: For 100 randomly chosen examples from the CIFAR-10 dataset, we show which rotations lead to a misclassification by various models trained with and without data augmentation ("rnd"). Each row corresponds to one example and each column to one angle. A dark red square indicates that the corresponding example was misclassified after being rotated by the corresponding angle. The visualization for ALP(mix, rnd) is more fragmented than for ALP(rob, rnd) and ALP(mix, rnd).

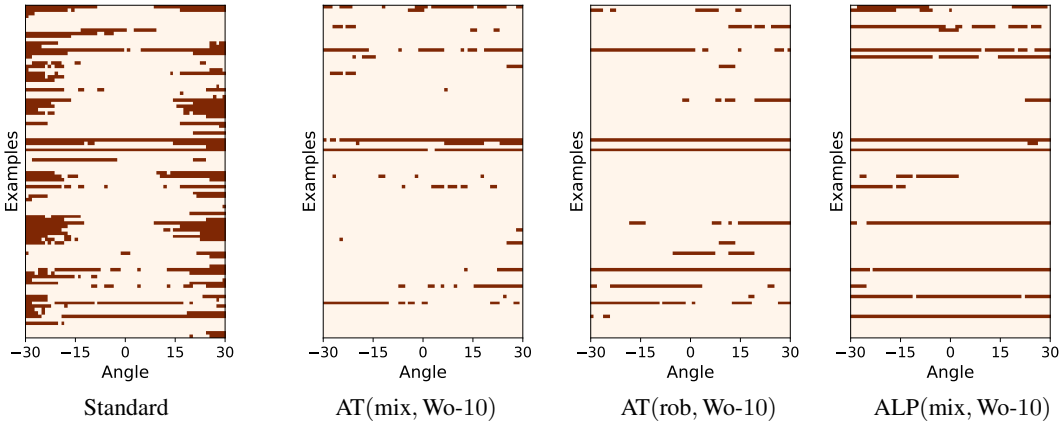

Figure 6: For 100 randomly chosen examples from the CIFAR-10 dataset, we show which rotations lead to a misclassification by various models trained with and without the Wo-10 defense. Each row corresponds to one example and each column to one angle. A dark red square indicates that the corresponding example was misclassified after being rotated by the corresponding angle. The visualization for AT(mix, Wo-10) is more fragmented than for AT(rob, Wo-10) and ALP(mix, Wo-10).