Do Perceptually Aligned Gradients Imply Adversarial Robustness?

Roy Ganz [1] **Bahjat Kawar** [2] **Michael Elad** [2]

Abstract

In the past decade, deep learning-based networks have achieved unprecedented success in numerous tasks, including image classification. Despite this remarkable achievement, recent studies have demonstrated that such networks are easily fooled by small malicious perturbations, also known as adversarial examples. This security weakness led to extensive research aimed at obtaining robust models. Beyond the clear robustness benefits of such models, it was also observed that their gradients with respect to the input align with human perception. Several works have identified Perceptually Aligned Gradients (PAG) as a byproduct of robust training, but none have considered it as a standalone phenomenon nor studied its own implications. In this work, we focus on this trait and test whether *Perceptually Aligned Gradients imply Robustness*. To this end, we develop a novel objective to directly promote PAG in training classifiers and examine whether models with such gradients are more robust to adversarial attacks. Extensive experiments on CIFAR-10 and STL validate that such models have improved robust performance, exposing the surprising bidirectional connection between PAG and robustness.

1. Introduction

AlexNet (Krizhevsky et al., 2012), one of the first Deep Neural Networks (DNNs), has significantly surpassed all the classic computer vision methods in the ImageNet (Deng et al., 2009) classification challenge (Russakovsky et al., 2015). Since then, the amount of interest and resources invested in the deep learning (DL) field has skyrocketed. Nowadays, such models attain superhuman performance in classification (He et al., 2015; Dosovitskiy et al., 2020). However, although neural networks are allegedly inspired by the human brain, unlike the human visual system, they are known to be highly sensitive to minor corruptions (Hosseini et al., 2017; Dodge & Karam, 2017; Geirhos et al., 2017; Temel et al., 2017; 2018; Temel & AlRegib, 2018) and small malicious perturbations, known as adversarial attacks (Szegedy et al., 2014; Athalye et al., 2017; Biggio et al., 2013; Carlini & Wagner, 2017b; Goodfellow et al., 2015; Kurakin et al., 2017; Nguyen et al., 2014). With the introduction of DNNs to real-world applications affecting human lives, these issues raise significant safety concerns, and therefore, have drawn substantial research attention.

The bulk of the works in the field of robustness to adversarial attacks can be divided into two types – on the one hand, ones that propose robustification methods (Goodfellow et al., 2015; Madry et al., 2018; Zhang et al., 2019; Wang et al., 2020), and on the other hand, ones that construct stronger and more challenging adversarial attacks (Goodfellow et al., 2015; Madry et al., 2018; Carlini & Wagner, 2017a; Tramer et al., 2020; Croce & Hein, 2020). While there are numerous techniques for obtaining adversarially robust models (Lecuyer et al., 2018; Li et al., 2018; Cohen et al., 2019b; Salman et al., 2019), the most effective one is Adversarial Training (AT) (Madry et al., 2018). AT proposes a simple yet highly beneficial training scheme – train the network to classify adversarial examples correctly. We provide an overview on adversarial examples and training in Appendix A.

While exploring the properties of adversarially trained models, Tsipras et al. (2019) exposed a fascinating characteristic of these models that does not exist in standard ones – Perceptually Aligned Gradients (PAG). Generally, they discovered that such models are more aligned with human perception than standard ones, in the sense that the loss gradients w.r.t. the input are meaningful and visually understood by humans. As a result, modifying an image to maximize a conditional probability of some class, estimated by a model with PAG, yields class-related semantic visual features, as can be seen in Figure 1. This important discovery has led to a sequence of works that uncovered conditions in which PAG occurs. Aggarwal et al. (2020) revealed that PAG also exists in adversarially trained models with small threat models, while Kaur et al. (2019) observed PAG in robust models trained without adversarial training. While it has been established that robust models lead to PAG, more research is required to better understand this intriguing property.

[1] Department of EE, Technion - Israel Institute of Technology, Haifa, Israel [2] Department of CS, Technion - Israel Institute of Technology, Haifa, Israel. Correspondence to: Roy Ganz <ganz.campus@technion.ac.il>.

In this work, while aiming to shed some light on the PAG phenomenon, we pose the following reversed question – *Do Perceptually Aligned Gradients Imply Robustness?* This is an interesting question, as it tests the similarity between neural networks and human vision. Humans are capable of identifying the class-related semantic features and thus, can describe the modifications that need to be done to an image to change their predictions. That, in turn, makes the human visual system "robust", as it is not affected by changes unrelated to the semantic features. With this insight, we hypothesize that since similar capabilities exist in classifiers with perceptually aligned gradients, they would be inherently more robust.

To methodologically test this question, we need to train networks that obtain perceptually aligned gradients without inheriting robust characteristics from robust models. However, PAG is known to be a byproduct of robust training, and there are currently no ways to promote this property directly and in isolation. Thus, to explore our research question, we develop a novel PAG-inducing general objective that penalizes the input-gradients of the classifier without any form of robust training. We first verify that our optimization goal indeed yields such gradients as well as sufficiently high accuracy on clean images, then test the robustness of the obtained models and compare them to models trained using standard training ("vanilla"). Our experiments strongly suggest that models with PAG are inherently more robust than their vanilla counterparts, revealing that directly promoting such a trait can imply robustness to adversarial attacks.

2. Perceptually Aligned Gradients

Perceptually aligned gradients (PAG) (Engstrom et al., 2019b; Etmann et al., 2019a; Ross & Doshi-Velez, 2018; Tsipras et al., 2019) is a phenomenon according to which, classifier input-gradients are semantically aligned with human perception. This means, inter alia, that modifying an image to maximize a specific class probability should yield visual features that humans associate with the target class. Tsipras et al. (2019) discovered that PAG occurs in adversarially trained classifiers, but not in "vanilla" models. The prevailing hypothesis is that the existence of PAG only in adversarially robust classifiers and not in regular ones indicates that features learned by such models are more aligned with human vision. PAG is a qualitative trait, and currently, no quantitative metrics for assessing it exist. Moreover, there is an infinite number of equally good gradients aligned with human perception, *i.e.*, there are countless perceptually meaningful directions in which one can modify an image to look more like a certain target class. Thus, in this work, similar to (Tsipras et al., 2019), we gauge PAG qualitatively by examining the visual modifications done while maximizing the conditional probability of some class, estimated by the tested classifier. In other words, we examine the effects of a large-ϵ targeted adversarial attack and claim that a model has PAG if such a process yields class-related semantic modifications, as demonstrated in Figure 1.

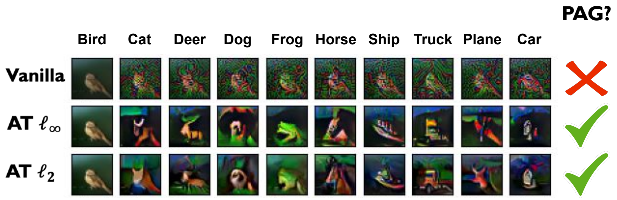


Figure 1. Visual demonstration of large-ϵ adversarial examples on "vanilla" and robust ResNet-18 trained on CIFAR-10 as a method to determine whether a model obtains PAG.

In recent years, PAG has drawn a lot of research attention which can be divided into two main types – an applicative study and a theoretical one. The applicative study aims to harness this phenomenon for various computer vision problems, such as image generation and translation (Santurkar et al., 2019), the improvement of state-of-the-art results in image generation (Ganz & Elad, 2021), and explainability (Elliott et al., 2021).

As for the theoretical study, several works aimed to better understand the conditions under which PAG occurs. The authors of (Kaur et al., 2019) examined if PAG is an artifact of the adversarial training algorithm or a general property of robust classifiers. Additionally, it has been shown that PAG exists in adversarially robust models with a low max-perturbation bound (Aggarwal et al., 2020). To conclude, previous works discovered that training robust models leads to PAG. In this work, we explore the opposite question – *Do perceptually aligned gradients imply robustness?*

3. Do PAG Imply Robustness?

As mentioned in Section 2, previous work has validated that robust training implies perceptually aligned gradients. More specifically, they observed that performing targeted PGD attacks on robust models yields visual modifications aligned with human perception. In contrast, in this work, we aim to delve into the opposite direction and test if training a classifier to have perceptually aligned gradients will improve its adversarial robustness.

To this end, we propose to encourage the input-gradients of a classifier f_θ to uphold PAG. Due to the nature of our research question, we need to isolate PAG from robust training and verify whether the former implies the latter. This raises a challenging question – PAG is known to be a byproduct of robust training. How can one develop a training procedure that encourages PAG without explicitly performing robust training of some sort? Note that a framework that attains

PAG via robust training cannot answer our question, as that would involve circular reasoning.

We answer this question by proposing a novel training objective consisting of two elements: the classic cross-entropy loss on the model outputs and an auxiliary loss on the model's input-gradients. We note that the input-gradients of the classifier, $\nabla_{\mathbf{x}} f_\theta(\mathbf{x})_y$, where $f_\theta(\mathbf{x})_y$ is the y-th entry of the vector $f_\theta(\mathbf{x})$, can be trained, since they are differentiable w.r.t. the classifier parameters θ. Thus, given labeled images $(\mathbf{x}, y)$ from a dataset D, assuming we have access to ground-truth perceptually aligned gradients $g(\mathbf{x}, y_t)$, we could pose the following loss function:

$$\begin{aligned}\mathcal{L}_{total}(\mathbf{x}, y) = \mathcal{L}_{CE}\left(f_\theta(\mathbf{x}), y\right) + \\ \lambda \sum_{y_t=1}^{C} \mathcal{L}_{cos}\left(\nabla_{\mathbf{x}} f_\theta(\mathbf{x})_{y_t}, g(\mathbf{x}, y_t)\right),\end{aligned} \tag{1}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss defined in Equation (4), λ is a tunable regularization hyperparameter, C is the number of classes in the dataset, and $\mathcal{L}_{cos}$ is the cosine similarity loss defined as follows:

$$\mathcal{L}_{cos}(\mathbf{v}, \mathbf{u}) = 1 - \frac{\mathbf{v}^\top \mathbf{u}}{\mathtt{max}(\|\mathbf{v}\|_2 \cdot \|\mathbf{u}\|_2, \varepsilon)}, \tag{2}$$

where ε is a small positive value so as to avoid division by zero. Note that this loss considers the *direction* of the model's input-gradients without any requirement on their magnitude. This bodes well with the general goal of these gradients being *aligned* with human perception.

We emphasize that, in contrast to robust training methods such as (Madry et al., 2018; Cohen et al., 2019a), our scheme does not feed the model with any perturbed images and only trains on examples originating from the training set. Moreover, while other works (Ross & Doshi-Velez, 2017; Jakubovitz & Giryes, 2018) suggest that penalizing the input-gradients' norm yields robustness, we do not utilize this fact since we encourage gradient alignment rather than having a small norm. Thus, our method is capable of promoting PAG without utilizing robust training.

After training a model to minimize the objective in Equation (1), we aim to examine if promoting PAG in a classifier increases adversarial robustness. First, to verify that the resulting model indeed upholds PAG, we perform targeted PGD on test set images and qualitatively assess the validity of the resulting visual modifications. Afterwards, we test the adversarial robustness of the said model and compare it with vanilla baselines. If it demonstrates favorable robustness accuracy, we will have promoted an affirmative answer to the titular research question of this work.

However, one major obstacle remains in the way of training this objective: so far, we have assumed the existence of "ground-truth" model input-gradients, an assumption that does not hold in practice. While we hypothesize that these gradients should point in the general direction of the target class images, there is no clear way of obtaining point-wise realizations of them. In the following section, we present practical and simple methods for obtaining approximations for these gradients, which we then use for training PAG-promoting classifiers.

4. How are "Ground Truth" PAGs Obtained?

In order to train a classifier for minimizing the objective in Equation (1), a "ground truth" perceptually aligned gradient $g(\mathbf{x}, y_t)$ needs to be provided for each training image $\mathbf{x} \in D$ and for each target class $y_t \in \{1, 2, \dots, C\}$. Since a true such gradient is infeasible to get, we instead explore two general pragmatic approaches for obtaining approximations for these PAGs.

4.1. Robust Input-Gradient Distillation

A possible realization of the "ground truth" perceptually aligned gradients may rely on the fact that adversarially trained models uphold PAG (Tsipras et al., 2019). One can use the input-gradients of such trained robust models and train a classifier to mimic them, according to Equation (1). More precisely, according to this realization, one can set $g(\mathbf{x}, y_t) = \nabla_{\mathbf{x}} f_\theta^*(\mathbf{x})_{y_t}$, where f_θ^* is an adversarially robust classifier. This way, we distill the PAG property of a robust classifier into our model. Similar ideas were explored in the context of robust knowledge distillation in (Chan et al., 2020; Shao et al., 2021; Sarkar et al., 2021). We discuss their connection and differences from our work in Appendix B.

At first glance, using such gradients in our objective seems different from adversarial training – while the latter relies on training the model to classify perturbed images correctly, our approach trains solely on clean images from the dataset, as in vanilla training. However, more careful inspection reveals that despite the clear benefits of this approach, it may lead to a leakage of the desired robustness properties to the trained model. This, in turn, makes testing the robustness of this model a form of circular reasoning, a logical fallacy. To better understand this crucial point, one can view this training approach as performing function approximation of f_θ^* via a first-degree Taylor series. More specifically, in the zeroth degree, we train the classifier to correctly classify training samples – an approximation of the outputs of the robust classifier. In the first degree approximation, we encourage the model to have similar gradients to the robust one. Thus, overall, since this technique pushes the trained model to mimic the behavior of the robust classifier, it cannot be utilized as a valid way of answering our titular question, and we use it only as an empirical performance upper bound. Thus, in order to ensure satisfactory conditions for assessing whether PAG implies robustness, we explore alternative sources of ground truth gradients that do not stem, explicitly nor implicitly, from adversarially trained models.

4.2. Target Class Representatives

As explained above, we aim to explore "ground truth" gradients that promote PAG without relying on robust models. To this end, we adopt the following simple premise: the gradient $g(\mathbf{x}, y_t)$ should point towards the general direction of images of the target class y_t. Therefore, given a representative of the target class, $\mathbf{r}_{y_t}$, we set the gradient to point away from the current image and towards the representative, *i.e.*, $g(\mathbf{x}, y_t) = \mathbf{r}_{y_t} - \mathbf{x}$. This general heuristic, visualized in Figure 2, can be manifested in various ways, of which we consider the following:

One Image: Each representative should be chosen to reflect the visual features of its respective class. The simplest such choice that comes to mind is to choose $\mathbf{r}_{y_t}$ as an arbitrary training set image with label y_t, and use it as a global destination of y_t-targeted gradients. This *one image* approach satisfies the abstract requirements and provides simplicity, but it introduces a strong bias towards the arbitrarily chosen representative image, without considering the target class as a whole.

Class Mean: In order to reduce the bias towards a single image, we can set $\mathbf{r}_{y_t}$ to be the mean of all the training images with label y_t. This mean can be multiplied by a constant in order to obtain an image-like norm. However, the *class mean* approach suffers from a clear limitation: a class' image distribution can be highly multimodal, possibly reducing its mean to a non-informative image.

Nearest Neighbor: As a possibly better trade-off, we examine a nearest neighbor (NN) approach – for each image $\mathbf{x}$ and each target class $y_t \in \{1, 2 \ldots, C\}$ we set the class representative $\mathbf{r}_{y_t}(\mathbf{x})$ (now dependent on the image) to be the image's NN amongst a limited set of samples from class y_t, using L_2 distance in the pixel space. More formally, we define

$$\mathbf{r}(\mathbf{x}, y_t) = \underset{\hat{\mathbf{x}} \in D_{y_t} \text{ s.t. } \hat{\mathbf{x}} \neq \mathbf{x}}{\operatorname{argmin}} \|\hat{\mathbf{x}} - \mathbf{x}\|_2, \tag{3}$$

where D_{y_t} is the set of sample images with class y_t. In practice, we sample D_{y_t} to be a small number of *i.i.d.* training set images with class y_t.

We test the aforementioned class representative approaches empirically in Section 5, owing to their relative simplicity and alignment with the abstract requirements on the gradients. However, we recognize that these approaches may oversimplify the desired behaviour, and may not be ideal. We therefore encourage the pursuit of more advanced options in future work. For example, the above two options (class mean and nearest neighbor) could be merged and extended by a preliminary clustering of each class subset to several sub-clusters, and an assignment of the NN amongst the obtained cluster means. Nevertheless, we choose not to explore this option further as it is more complex, losing much of the appeal and the simplicity of the options described above.

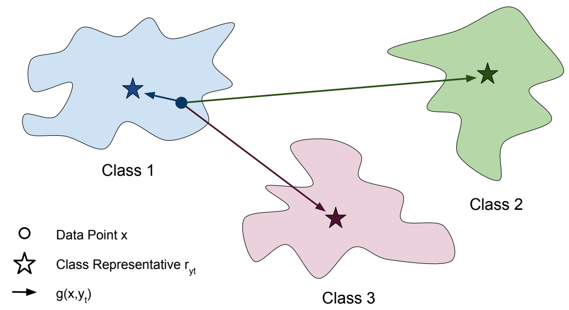


Figure 2. An illustration of the proposed creation of perceptually meaningful gradients for the training process as given in Eq. 1.

5. Experimental Results

In this section we empirically assess whether promoting PAG during classifier training improves its adversarial robustness at test time. We experiment using both synthetic and real datasets and present our findings in Section 5.1 and Section 5.2, respectively.

5.1. A Toy Dataset

To illustrate and better understand the proposed approach and its effects, we experiment with a synthetic 2-dimensional dataset and compare our *nearest neighbor* approach with the vanilla training scheme that minimizes the cross-entropy loss. We construct a dataset consisting of 6,000 samples from two classes, where each class contains exactly 3,000 examples, and the manifold assumption upholds – the data resides on a lower-dimensional manifold (Ruderman, 1994). We train a classifier (two-layer fully-connected network) twice: once with our nearest neighbor method, and once without it. We then examine the obtained accuracies and visualize the decision boundaries. While both methods reach a perfect accuracy over the test set, the obtained decision boundaries differ substantially, as can be seen in Figure 3. The baseline training method results in Figure 3a yields dimpled manifolds as decision boundaries, as hypothesized in (Shamir et al., 2021). According to their hypothesis, since the decision boundary of DNN is very close to the data manifold, adversarial examples very close to the image manifold exist, exposing the model to malicious perturbations.

In contrast, as can be seen in Figure 3b, the margin between the data samples and the decision boundary obtained using our approach is significantly larger than the baseline. This observation helps explain the following robustness result: our model achieves a 75.5% accuracy on a simple adversarial PGD attack, whereas the baseline model collapses to 0.0%. We provide additional details regarding this experiment in the supplementary material. Note that the notion of "perceptually aligned" gradients admits a very clear meaning in the context of our 2-dimensional experiment – faithfulness to the known data manifold. In the baseline approach, the gradients used in adversarial attacks will point towards close areas of the opposite class, orthogonal to the data manifold. Thus, such gradients deviate from the data behavior and are not aligned with its distribution. In contrast, in our approach, due to the absence of close orthogonal misclassified areas, adversarial gradients tend to be more aligned with the data manifold, making it perceptually meaningful.

5.2. Experimenting with Real Datasets

With the encouraging findings presented in Section 5.1, we now turn to conduct thorough experiments to verify if indeed promoting PAG can lead to improved adversarial robustness on real datasets – CIFAR-10 (Krizhevsky et al., 2014) and STL (Coates et al., 2011). In order to make a well-founded empirical claim, we explore the several sources for "ground truth" PAG proposed in Section 4.1 and Section 4.2.

To verify if promoting PAG implies robustness, we first validate that such a phenomenon (i.e., PAG) occurs when using our method, and then we test the performance of such models under attacks. We start by training a classifier and qualitatively examine whether our approach leads to perceptually aligned gradients. More specifically, we probe whether modifying an image to maximize a certain class probability, estimated by a model, leads to a meaningful semantic change. Then we turn to assess the performance of our method using two main metrics – clean accuracy and adversarial robustness using both the ℓ_∞ and the ℓ_2 threat models. More specifically, we use a 20-step PGD (PGD^{20}) as our adversarial attack, with $\epsilon = 0.031$ for ℓ_∞ attacks on both datasets, and $\epsilon = 0.5,\ 1.5$ for ℓ_2 on CIFAR-10 and STL, respectively. We include additional implementation details regarding the attack in the supplementary material. If we find that our method attains improved robust accuracy compared to standard training, we provide evidence that promoting PAG can improve robustness.

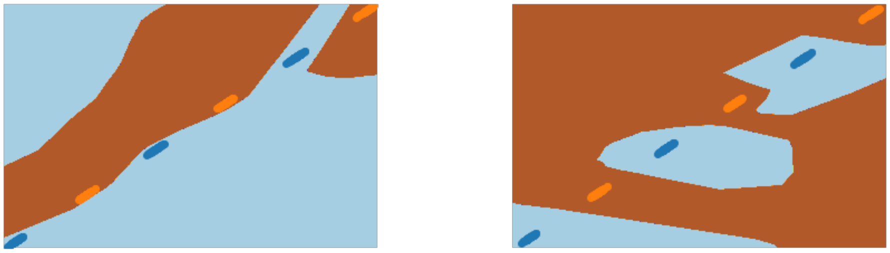
In all the conducted experiments, we train a ResNet-18 (He et al., 2015) classifier to minimize Equation (1). To validate if indeed promoting PAG implies adversarial robustness, we compare it with a "vanilla" training using the same hyperparameters (Additional implementation details are listed in the supplementary material). We train the model using two main ground-truth gradients sources listed in Section 4 – target class representatives and Robust Input-Gradient Distillatios (RIGD). While the methods of the first, One Image (OI), Class Mean (CM), and Nearest Neighbor (NN), are valid options for verifying the research question, RIGD serves only as an empirical upper bound due to the logical fallacy detailed in Section 4.1.

Results: As explained above, we determine that a method promotes PAG if the pixel-space modifications done while maximizing the conditional probability induced by a classifier align with human perception. We show in Figure 4 that while vanilla models do not exhibit semantically meaningful changes, our approach does. In addition, the modifications obtained by our method are similar to the ones of adversarially trained models, visualized in Figure 1, and both contain rich class-related information. Note that in our method, we only train the model to have point-wise gradients aligned with some "ground truth" ones. However, surprisingly, it is able to guide the iterative maximization process towards semantically meaningful modifications, although never trained on these intermediate points. In other words, although trained to have aligned gradients to some ground truth ones only on the data points, the model generalizes to have meaningful gradients far beyond these points.

We proceed by quantitatively evaluating the performance on clean and adversarial versions of the test set, as mentioned above, and show our results in Tables 1 and 2. All the tested PAG-inducing techniques improve the adversarial robustness substantially, while maintaining competitive clean accuracies. While the vanilla baseline is utterly vulnerable to adversarial examples, introducing a PAG-inducing auxiliary objective robustifies it without performing robust training. This surprising finding strongly suggests that promoting PAG can improve the classifier's robustness. As the results indicate, our method performs better in the L_2 case over the L_∞ one. We hypothesize that this stems from the Euclidean nature of the cosine similarity loss used to penalize the model gradients. We emphasize that while superior robustification methods rely on training to correctly classify (adversarially) perturbed images, our scheme achieves significant robustness, solely by promoting PAG.

6. Conclusion and Future Work

While previous work demonstrate that adversarially robust models uphold the Perceptually Aligned Gradients property, in this work, we aim to investigate the reverse question – *Do Perceptually Aligned Gradients Imply Adversarial Robustness?* We believe that answering this question sheds additional light on the connection between robust models and PAG. To empirically show that inducing PAG improves classifier robustness, we develop a novel generic optimization loss for promoting PAG without relying on robust models or adversarial training, and test several manifestations of it. Our findings suggest that all such methods that pro-



(a) Vanilla Training Scheme

(b) Our Training Scheme

Figure 3. Visualization of the decision boundary on a synthetic two-class dataset – the points are the test samples, and the background color represents the predicted class. Figures 3a and 3b present the decision boundary of a vanilla training method and ours, respectively. Our approach increases the margin between the instances and the decision boundary, yielding improved robustness.

Table 1. Accuracy scores on the CIFAR-10 dataset using the ResNet-18 architecture.

Method	No Attack	L_∞	L_2
Vanilla	93.61%	00.00%	00.00%
One Image	79.46%	16.57%	49.68%
Class Mean	81.41%	15.65%	50.62%
Nearest Neighbor	80.65%	10.58%	46.33%
RIGD ℓ_∞	84.94%	40.69%	58.51%
RIGD ℓ_2	88.96%	28.35%	65.04%

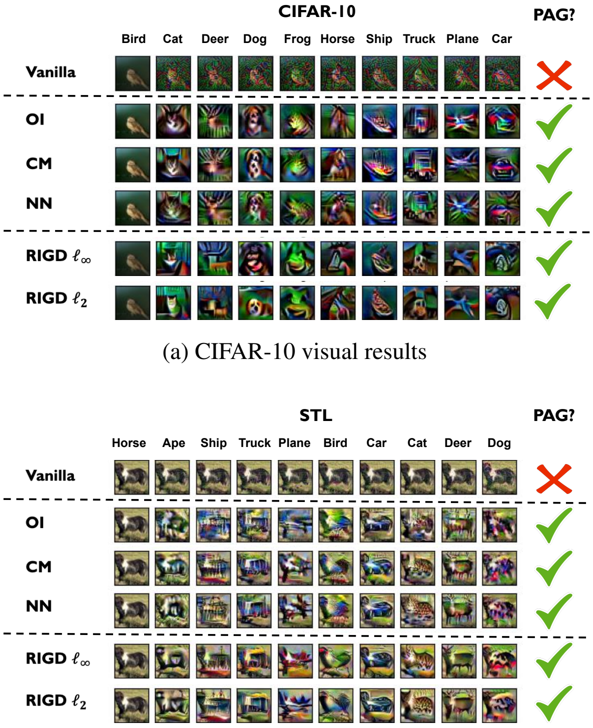
Table 2. Accuracy scores on the STL dataset using the ResNet-18 architecture.

Method	No Attack	L_∞	L_2
Vanilla	82.60%	00.00%	00.00%
One Image	71.30%	10.65%	36.65%
Class Mean	70.64%	14.58%	40.65%
Nearest Neighbor	70.16%	16.35%	41.94%
RIGD ℓ_∞	75.19%	26.13%	42.65%
RIGD ℓ_2	75.18%	20.66%	48.46%

mote PAG improve the adversarial robustness compared to a vanilla model, while still maintaining an adequate clean accuracy. Despite that, the obtained robustness still falls behind state-of-the-art models, which possibly stems from oversimplified realizations of "ground-truth" PAGs. Therefore, we believe that improving these realizations would be a key factor for the continued success of the proposed training objective in future work.

7. Acknowledgements

This research was partially supported by the Israel Science Foundation (ISF) under Grant 335/18, the Israeli Council for Higher Education – Planning & Budgeting Committee, and the Stephen A. Kreynes Fellowship.



(a) CIFAR-10 visual results

(b) STL visual results

Figure 4. Perceptually Aligned Gradients phenomenon demonstrated by models trained with vanilla training (top), our method (middle), and our gradient distillation baseline (bottom), all using ResNet-18 on the CIFAR-10 and STL datasets.

References

Aggarwal, G., Sinha, A., Kumari, N., and Singh, M. K. On the benefits of models with perceptually-aligned gradients. *ArXiv*, abs/2005.01499, 2020.

Andriushchenko, M. and Flammarion, N. Understanding and improving fast adversarial training. In *NeurIPS*, 2020.

Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. Synthesizing robust adversarial examples, 2017.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. *Lecture Notes in Computer Science*, pp. 387–402, 2013. ISSN 1611-3349.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017a.

Carlini, N. and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods, 2017b.

Chan, A., Tay, Y., and Ong, Y.-S. What it thinks is important is important: Robustness transfers through input gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 332–341, 2020.

Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In Gordon, G., Dunson, D., and Dudík, M. (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 215–223, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320. PMLR, 09–15 Jun 2019a.

Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. Certified adversarial robustness via randomized smoothing, 2019b.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks, 2020.

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

Dodge, S. and Karam, L. A study and comparison of human and deep learning recognition performance under visual distortions, 2017.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., and Li, J. Boosting adversarial attacks with momentum. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

Elliott, A., Law, S., and Russell, C. Explaining classifiers using adversarial perturbations on the perceptual ball. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Engstrom, L., Ilyas, A., Salman, H., Santurkar, S., and Tsipras, D. Robustness (python library), 2019a. URL https://github.com/MadryLab/robustness.

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Tran, B., and Madry, A. Adversarial robustness as a prior for learned representations. *arXiv: Machine Learning*, 2019b.

Etmann, C., Lunz, S., Maass, P., and Schoenlieb, C. On the connection between adversarial robustness and saliency map interpretability. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1823–1832. PMLR, 09–15 Jun 2019a.

Etmann, C., Lunz, S., Maass, P., and Schoenlieb, C. On the connection between adversarial robustness and saliency map interpretability. In *International Conference on Machine Learning*, pp. 1823–1832. PMLR, 2019b.

Finlay, C. and Oberman, A. M. Scaleable input gradient regularization for adversarial robustness. *Machine Learning with Applications*, 3:100017, 2021.

Ganz, R. and Elad, M. Bigroc: Boosting image generation via a robust classifier. *CoRR*, abs/2108.03702, 2021.

Geirhos, R., Janssen, D. H. J., Schütt, H. H., Rauber, J., Bethge, M., and Wichmann, F. A. Comparing deep neural networks against humans: object recognition when the signal gets weaker, 2017.

Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations, ICLR*, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015.

Hosseini, H., Xiao, B., and Poovendran, R. Google's cloud vision api is not robust to noise, 2017.

Huang, L., Zhang, C., and Zhang, H. Self-adaptive training: beyond empirical risk minimization. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Jakubovitz, D. and Giryes, R. Improving dnn robustness to adversarial attacks using jacobian regularization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 514–529, 2018.

Kaur, S., Cohen, J., and Lipton, Z. Are perceptually-aligned gradients a general property of robust classifiers?, 10 2019.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pp. 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

Krizhevsky, A., Nair, V., and Hinton, G. The cifar-10 dataset. *online: http://www. cs. toronto. edu/kriz/cifar. html*, 55 (5), 2014.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial examples in the physical world, 2017.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with differential privacy, 2018.

Li, B., Chen, C., Wang, W., and Carin, L. Certified adversarial robustness with additive noise, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Nguyen, A. M., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *CoRR*, abs/1412.1897, 2014.

Pang, T., Yang, X., Dong, Y., Xu, T., Zhu, J., and Su, H. Boosting adversarial training with hypersphere embedding. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Qin, C., Martens, J., Gowal, S., Krishnan, D., Dvijotham, K., Fawzi, A., De, S., Stanforth, R., and Kohli, P. Adversarial robustness through local linearization. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 13824–13833, 2019.

Ross, A. and Doshi-Velez, F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Ross, A. S. and Doshi-Velez, F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients, 2017.

Ross, A. S. and Doshi-Velez, F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In McIlraith, S. A. and Weinberger, K. Q. (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 1660–1669. AAAI Press, 2018.

Ruderman, D. L. The statistics of natural images. *Network: Computation in Neural Systems*, 5(4):517–548, 1994.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

Salman, H., Yang, G., Li, J., Zhang, P., Zhang, H., Razenshteyn, I., and Bubeck, S. Provably robust deep learning via adversarially trained smoothed classifiers, 2019.

Santurkar, S., Tsipras, D., Tran, B., Ilyas, A., Engstrom, L., and Mądry, A. *Image Synthesis with a Single (Robust) Classifier*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Sarkar, A., Sarkar, A., Gali, S., and N Balasubramanian, V. Get fooled for the right reason: Improving adversarial robustness through a teacher-guided curriculum learning approach. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12836–12848, 2021.

Shamir, A., Melamed, O., and BenShmuel, O. The dimpled manifold model of adversarial examples in machine learning, 2021.

Shao, R., Yi, J., Chen, P.-Y., and Hsieh, C.-J. How and when adversarial robustness transfers in knowledge distillation? *arXiv preprint arXiv:2110.12072*, 2021.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations, ICLR*, 2014.

Temel, D. and AlRegib, G. Traffic signs in the wild: Highlights from the IEEE video and image processing cup 2017 student competition [SP competitions]. *IEEE Signal Processing Magazine*, 35(2):154–161, mar 2018.

Temel, D., Kwon, G., Prabhushankar, M., and AlRegib, G. Cure-tsr: Challenging unreal and real environments for traffic sign recognition. 2017.

Temel, D., Lee, J., and AlRegib, G. Cure-or: Challenging unreal and real environments for object recognition. 2018.

Tramer, F., Carlini, N., Brendel, W., and Madry, A. On adaptive attacks to adversarial example defenses, 2020.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness May Be at Odds with Accuracy. In *International Conference on Learning Representations, ICLR*, 2019.

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., and Gu, Q. Improving adversarial robustness requires revisiting misclassified examples. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Xie, C., Wu, Y., van der Maaten, L., Yuille, A. L., and He, K. Feature denoising for improving adversarial robustness. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 501–509. Computer Vision Foundation / IEEE, 2019.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7472–7482. PMLR, 2019.

A. Background

A.1. Adversarial Examples

We consider a deep learning-based classifier $f_\theta : \mathbb{R}^M \to \mathbb{R}^C$, where M is the data dimension and C is the number of classes. Adversarial examples are instances designed by an adversary in order to cause a false prediction by f_θ (Athalye et al., 2017; Biggio et al., 2013; Carlini & Wagner, 2017b; Goodfellow et al., 2015; Kurakin et al., 2017; Nguyen et al., 2014; Szegedy et al., 2014). In 2013, Szegedy et al. (2014) discovered the existence of such samples and showed that it is possible to cause misclassification of an image with an imperceptible perturbation, which is obtained by maximizing the network's prediction error. Such samples are crafted by applying modifications from a *threat model* Δ to real natural images. Hypothetically, the "ideal" threat model should include all the possible label-preserving perturbations, *i.e.*, all the modifications that can be done to an image that will not change a human observer's prediction. Unfortunately, it is impossible to rigorously define such Δ, and thus, simple relaxations of it are used, the most common of which are the ℓ_2 and the ℓ_∞ ϵ-balls: $\Delta = \{\delta \ : \ \|\delta\|_{c \in \{2,\infty\}} \leq \epsilon\}$.

More formally, given an input sample $\mathbf{x}$, its ground-truth label y and a threat model Δ, a valid adversarial example $\hat{\mathbf{x}}$ satisfies the following: $\hat{\mathbf{x}} = \mathbf{x} + \delta \ s.t. \ \delta \in \Delta, y_{pred} \neq y$, where y_{pred} is the prediction of the classifier on $\hat{\mathbf{x}}$. The procedure of obtaining such examples is referred to as an *adversarial attack*. Such attacks can be either untargeted or targeted. Untargeted attacks generate $\hat{\mathbf{x}}$ to minimize $p_\theta(y|\hat{\mathbf{x}})$, namely, cause a misclassification without a specific target class. In contrast, targeted attacks aim to craft $\hat{\mathbf{x}}$ in a way that maximizes $p_\theta(\hat{y}|\hat{\mathbf{x}}) \ s.t. \ \hat{y} \neq y$, that is to say, fool the classifier to predict $\hat{\mathbf{x}}$ as a target class $\hat{y}$.

While there are various techniques for generating adversarial examples (Goodfellow et al., 2015; Carlini & Wagner, 2017a; Dong et al., 2018), we focus in this work on the Projected Gradient Descent (PGD) method (Madry et al., 2018). PGD is an iterative procedure for obtaining adversarial examples that operates as described in Algorithm 1. The operation $Proj_\epsilon$ stands for a projection operator onto Δ, and $\mathcal{L}(\cdot)$ is the classification loss, usually defined as the cross-entropy:

$$\mathcal{L}_{CE}(\mathbf{z}, y) = -\log \frac{\exp(\mathbf{z}_y)}{\sum_{i=1}^{C} \exp(\mathbf{z}_i)}, \tag{4}$$

where $\mathbf{z}_y$ and $\mathbf{z}_i$ are the classifier's logits for classes y and i, respectively.

Algorithm 1 Projected Gradient Descent

```
Input: classifier $f_\theta$, input $\mathbf{x}$, ground-truth class $y$, target class $\hat{y}$, threat model parameter $\epsilon$, step size $\alpha$, number of iterations $T$
$\delta_0 \leftarrow 0$
for t from 0 to T do
    if $\hat{y}$ is not None then
        $\delta_{t+1} = Proj_\epsilon(\delta_t - \alpha \nabla_\delta \mathcal{L}(f_\theta(\mathbf{x} + \delta_t), \hat{y}))$
    else
        $\delta_{t+1} = Proj_\epsilon(\delta_t + \alpha \nabla_\delta \mathcal{L}(f_\theta(\mathbf{x} + \delta_t), y))$
    end
end
$\mathbf{x}_{adv} = \mathbf{x} + \delta_T$
Output: $\mathbf{x}_{adv}$
```

A.2. Adversarial Training

Adversarial training (AT) (Madry et al., 2018) is a learning procedure that aims to obtain adversarially robust classifiers. A classifier is considered adversarially robust if applying small adversarial perturbations to its input does not change its label prediction (Goodfellow et al., 2015). Such classifiers can be obtained by solving the following optimization problem:

$$\min_\theta \sum_{(\mathbf{x},y) \in D} \max_{\delta \in \Delta} \mathcal{L}(f_\theta(\mathbf{x} + \delta), y). \tag{5}$$

Intuitively, the above optimization trains the classifier to accurately predict the class labels of its hardest perturbed images allowed by the threat model Δ. Ideally, $\mathcal{L}$ is the 0-1 loss, *i.e.*, $\mathcal{L}(\mathbf{z}, y) = \mathbf{I}(\texttt{argmax}_i(\mathbf{z}_i) = y)$ where $\mathbf{I}$ is the indicator function. Nevertheless, since the 0-1 loss is not differentiable, the cross-entropy loss, defined in Equation (4), is used as a surrogate. In practice, solving this min-max optimization problem is challenging, and there are several ways to obtain an approximate solution. The most simple yet effective method is based on approximating the solution of the inner-maximization via adversarial attacks, such as PGD (Madry et al., 2018). According to this strategy, the above optimization is performed iteratively by first fixing the classifier's parameters θ and optimizing the perturbation δ for each example via PGD and then fixing δ and updating θ. Repeating these steps results in a robust classifier. Since its introduction by Madry et al. (2018), various improvements to adversarial training were proposed (Andriushchenko & Flammarion, 2020; Huang et al., 2020; Pang et al., 2020; Qin et al., 2019; Xie et al., 2019; Zhang et al., 2019; Wang et al., 2020), yet in this work we will focus mainly on the basic AT scheme (Madry et al., 2018) for its simplicity and generality.

B. Related Work

In Section 4.1 we demonstrate how robust input-gradient distillation can promote adversarial robustness in the trained model. This phenomenon has also been observed and harnessed in the context of Robust Knowledge Distillation by several recent papers (Chan et al., 2020; Shao et al., 2021; Sarkar et al., 2021). In these works, a *student* classifier model is trained to have similar gradients to a robust *teacher* model. The similarity is defined as either cosine similarity or the inability of a discriminator network to distinguish between the gradients of the two models. These works differ from our work, both in their loss functions and specific utilization of the teacher model, but essentially, they all demonstrate how distilling knowledge from a robust teacher model can invoke adversarial robustness in the student model. While successful in their respective tasks, these methods are not suitable for assessing whether perceptually-aligned gradients inherently promote robustness, as they are implicitly reliant on prior adversarial training, as explained in Section 4.1.

In addition, recent works have explored properties of input-gradients that improve adversarial robustness. The authors of (Jakubovitz & Giryes, 2018) demonstrate that regularizing the Frobenius norm of a classifier's Jacobian to be small, improves robustness. Such a method is equivalent to regularizing the norm of each such gradient to be small, similar to (Ross & Doshi-Velez, 2018; Finlay & Oberman, 2021). The work in (Etmann et al., 2019b) follows suit and considers the cosine similarity between a classifier's input-gradient w.r.t. the ground truth class and the input image itself. A positive correlation between this similarity and the classifier's adversarial robustness is observed. Nevertheless, none of these works promotes nor exhibits perceptually aligned gradients. In contrast, our work proposes to test the relation between the alignment of gradients with human perception and adversarial robustness and presents several PAG-promoting methods, inducing improved robustness.

C. Implementation Details

C.1. Toy Dataset

Data: We experiment with our approach on a 2-dimensional synthetic dataset to demonstrate its effects. As explained in the corresponding section in the paper, we construct a 2-class dataset where the data points reside on a straight line. Each class contains three modes, and each of them contains 1000 samples drawn from a Gaussian distribution ($x_1 \sim N(c, 1), x_2 = 2 * x_1$, where c is the mode center). The modes centers are set to be $\{-50, -10, 30\}$ and $\{-30, 10, 50\}$.

Architecture and Training: We use a 2-layer fully-connected network $(2 \to 32 \to 2)$ with ReLU non-linearity. We train it twice – using the standard cross-entropy training and our proposed method with NN realization. We do so for 100 epochs with a batch size of 128, using Adam optimizer, a learning rate of 0.01, and the same seed for both training processes.

Computational Resources: We use a single GPU via the Google Colab service.

Evaluation: As detailed in the paper, we test the performance of the models using clean and adversarial evaluation. For the clean one, we draw 600 test samples from the same distribution as the train set and measure the accuracy. As for the adversarial one, we use an L_2-based 10-step PGD with $\epsilon = 15$ and a step size of 2. Note that this choice of ϵ guarantees in our settings that the allowed threat model is too small for actually changing a sample of a certain class to the other one, making it a valid threat model.

C.2. Real Datasets

Data: As for our real datasets experiments, we use CIFAR-10 and STL that contain images of size $32 \times 32 \times 3$ and $96 \times 96 \times 3$, respectively. For each realization, before the training procedure, we construct a dataset by computing C targeted gradients for each training sample ($C = 10$ for CIFAR-10 and STL) for reproducibility and consistency purposes. While our target class representatives methods are model-free, RIGD requires a robust classifier to distill its targeted gradients. For CIFAR-10, we use publicly available ResNet-50 (Engstrom et al., 2019a), trained on L_∞ and L_2 attacks with $\epsilon = 8/255$ and $\epsilon = 0.5$, respectively. For STL, due to the lack of pretrained models, we adversarially train two ResNet-18 classifiers, using L_∞ and L_2 threat models with $\epsilon = 8/255$ and $\epsilon = 1.5$, respectively.

Training: For bothdatasets, we train a ResNet-18 for 100 epochs, using SGD with a learning rate of 0.01, a momentum of 0.9, and a weight decay of 0.0001. In addition, we use the standard augmentations for these datasets – random cropping with padding of 4 and random horizontal flipping with a probability of 0.5. We use a batch size of 64 for CIFAR-10 and 32 for STL. We present in Table 3 the best choices of λ – the coefficient of our PAG promoting auxiliary loss term in all the tested datasets and methods.

Table 3. Values of hyperparameter λ.

Method	λ value	
	CIFAR-10	STL
One Image	0.5	0.25
Class Mean	0.4	0.2
Nearest Neighbor	0.4	0.4
RIGD ℓ_∞	0.4	0.4
RIGD ℓ_2	0.4	0.4

Computational Resources: We use a single NVIDIA A40 GPU for each experiment.

Evaluation: We use TRADES[1] code repository for adversarial attacks and extend it to contain L_2 attack, in addition to the existing L_∞ one. For assessing the adversarial robustness, we use a k-step PGD ($k = 20$), with random initialization and a step size of $\frac{2\epsilon}{k}$.

To validate that our training method is stable and consistent, we run CIFAR-10 experiments using the One Image method three times using different seeds and report the results in the Table 4 below. As the results indicate, our approach consistently leads to improved robustness.

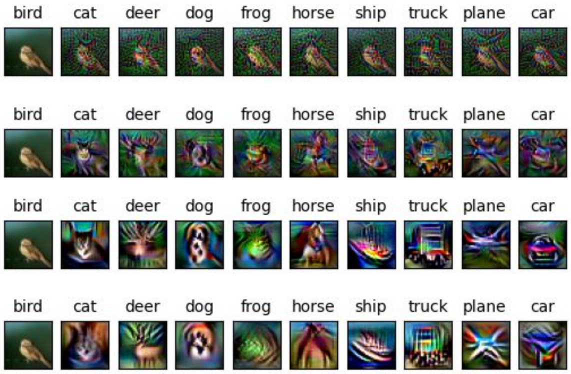
Table 4. Error bar evaluation on CIFAR-10 using One Image.

No Attack	PGD^{20}, L_∞	PGD^{20}, L_2
79.13 ± 0.74	15.93 ± 0.58	50.26 ± 0.91

D. Ablation Study

In this section, we test the effect of choosing different values of λ, *i.e.*, the coefficient of the PAG promoting auxiliary loss, on the CIFAR-10 dataset, using the Class Mean method. λ is a crucial hyperparameter as it trades off between clean accuracy and the PAG property and thus, changes the level of PAG, robust and clean accuracies. We summarize the results of this ablation in Figure 5. As can be seen, in general, increasing λ leads to more robust models with gradients better aligned with human perception. However, it comes with the cost of accuracy. We hypothesize that more sophisticated realizations of "ground-truth" PAG gradients can mitigate the tradeoff between accuracy and PAG.

[1] https://github.com/yaodongyu/TRADES



λ	Clean	ℓ_∞	ℓ_2
0.01	93.39%	0%	0.65%
0.1	88.40%	4.88%	36.43%
0.4	81.41%	15.65%	50.62%
1	72.63%	19.47%	48.73%

Figure 5. Quantitative and qualitative results of different λ values on CIFAR-10 using Class Mean.