Do Perceptually Aligned Gradients Imply Robustness?

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

Abstract

Adversarially robust classifiers possess a trait that non-robust models do not – Perceptually Aligned Gradients (PAG). Their gradients with respect to the input align well with human perception. Several works have identified 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 multiple datasets and architectures validate that models with aligned gradients exhibit significant robustness, exposing the surprising bidirectional connection between PAG and robustness. Lastly, we show that better gradient alignment leads to increased robustness and harness this observation to boost the robustness of existing adversarial training techniques. Our code is available at `https://github.com/royg27/PAG-ROB`.

1. Introduction

Since the tremendous success of AlexNet (Krizhevsky et al., 2012), one of the first Deep Neural Networks (DNNs), in the ImageNet (Deng et al., 2009) classification challenge, 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., 2016; Dosovitskiy et al., 2021). However, although neural networks are 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., 2018; Biggio et al., 2013; Carlini & Wagner, 2017b; Goodfellow et al., 2015; Kurakin et al., 2017; Nguyen et al., 2015). With the introduction of such models to real-world applications that affect human lives, these issues raise significant safety concerns, and therefore, they have drawn substantial research attention.

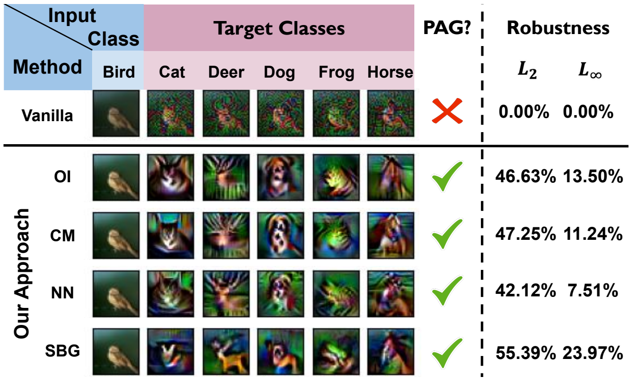


Figure 1. **PAG implies robustness.** Our method disentangles PAG from adversarial training and allows for a non-adversarial PAG-inducing training scheme that leads to substantial robustness.

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; Tramèr et al., 2020; Croce & Hein, 2020b). While there are numerous techniques for obtaining adversarially robust models (Lécuyer et al., 2019; Li et al., 2019; Cohen et al., 2019b; Salman et al., 2019), the most effective one is Adversarial Training (AT) (Madry et al., 2018) and its variants (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). AT proposes a simple yet highly beneficial training scheme – train the network to classify adversarial examples correctly.

While exploring the properties of adversarially trained models, Tsipras et al. (2019) exposed a fascinating characteristic

[1] Electrical Engineering Department, Technion, Haifa, Israel [2] Computer Science Department, Technion, Haifa, Israel. Correspondence to: Roy Ganz <ganz@campus.technion.ac.il>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

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 2. 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 different variations of robust models lead to perceptually aligned gradients, more research is required to understand this intriguing property better.

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.

Methodologically testing this question requires training classifiers to obtain PAG without performing robust training. However, this is challenging as 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. In particular, our objective encourages the classifier's gradients to be aligned with "ground-truth" ones that possess PAG. As such, our objective requires access to "ground-truth" such gradients which are challenging to obtain. Thus, we explore both heuristic and principled sources for these gradients. Our heuristic sources stem from the rationale that PAG should point toward the target class. In addition, we provide in this work a second, principled approach towards creating such PAG vectors, relying on denoising score matching as used in diffusion models (Song & Ermon, 2019). Specifically, we develop a theoretically justified approach to extract gradients that uphold PAG from diffusion models and encourage the gradients of the classifier to be aligned with ones distilled from diffusion models.

To validate our hypothesis, we first verify that our optimization goal indeed yields PAG and sufficiently high accuracy on clean images, then evaluate 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. We examine this implication across multiple datasets and architectures and conclude that it generally holds and is not data or model dependant. Surprisingly, although our primary goal is to shed light upon the connection between PAG and robustness rather than proposing a robustification method, not only does our method yield models with non-trivial robustness, but it also exhibits comparable robustness performance to adversarial training without training on perturbed images. Interestingly, promoting the gradient alignment of a model using our method outperforms adversarial training in the low-data regime and vision-transformers when trained from scratch. Moreover, it demonstrates better generalization to large ϵ attacks than AT. Thus, our findings can potentially pave the way for standard training methods (*i.e.*, without performing adversarial training) for obtaining robust classifiers. In addition, we study whether there is a correlation between the level of gradient alignment and robustness and discover that models with more aligned gradients are more robust to adversarial examples. Lastly, we harness this insight and introduce our PAG-inducing objective as an auxiliary loss in standard adversarial training techniques and observe that it significantly improves their robustness. In particular, we show that improving the gradient alignment of classifiers trained with AT (Madry et al., 2018) and TRADES (Zhang et al., 2019) improves their robustness to seen and unseen attacks by up to 2.24% and 5.25%, respectively. To summarize:

- We propose a methodological approach to train classifiers to possess PAG without performing adversarial training.
- We show that models with aligned gradients are inherently more robust, exposing the bidirectional connection between PAG and robustness.
- We demonstrate that increasing gradient alignment improves robustness and leverage this observation to improve existing robustification methods.

2. Background

2.1. 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., 2018; Biggio et al., 2013; Carlini & Wagner, 2017b; Goodfellow et al., 2015; Kurakin et al., 2017; Nguyen et al., 2015; Szegedy et al., 2014). 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: $\Delta = \{\delta \,:\, \|\delta\|_{c\in\{2,\infty\}} \le \epsilon\}$.

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 in iterative manner as described in Equation (1) below:

$$Repeat : \delta = Proj_\epsilon(\delta + \alpha \nabla_\delta \mathcal{L}(f_\theta(\mathbf{x} + \delta_t), y)). \quad (1)$$

$Proj_\epsilon$ is a projection operator onto Δ, α is the step size, and $\mathcal{L}(\cdot)$ is the classification loss, usually the cross-entropy:

$$\mathcal{L}_{CE}(\mathbf{z}, y) = -\log \frac{\exp(\mathbf{z}_y)}{\sum_{i=1}^{C} \exp(\mathbf{z}_i)}, \quad (2)$$

where $\mathbf{z}_y$ and $\mathbf{z}_i$ are the y and i logits, respectively. In addition, PGD can be extended to targeted attacks.

2.2. Adversarial Training

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

$$\min_\theta \sum_{(\mathbf{x},y)\in D} \max_{\delta\in\Delta} \mathcal{L}(f_\theta(\mathbf{x} + \delta), y). \quad (3)$$

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 (2), 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), yielding classifiers with improved robustness.

2.3. Perceptually Aligned Gradients

Perceptually aligned gradients (PAG) (Engstrom et al., 2019; Etmann et al., 2019; Ross & Doshi-Velez, 2018a; Tsipras et al., 2019) is a phenomenon according to which classifier input-gradients are semantically aligned with human perception. It 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" ones. 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, there are no quantitative metrics for assessing it. 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, as estimated by the tested classifier. In other words, we examine the effects of a large-ϵ targeted adversarial attack and determine that a model has PAG if such a process yields class-related semantic modifications, as demonstrated in Figure 2. As can be seen, adversarially trained models obtain PAG, while standardly trained ones ("vanilla") do not.

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

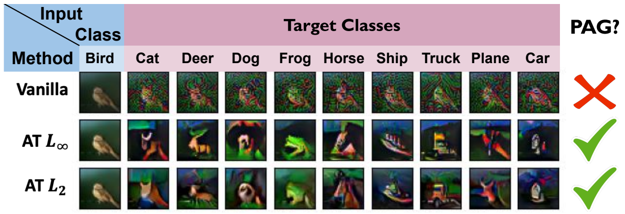


Figure 2. **Demonstration of PAG property**. Targeted large-ϵ adversarial examples on robust ResNet-18 trained on CIFAR-10 contain class-related modifications, while "vanilla" ones do not.

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 understand better the conditions under which PAG occurs. 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 robust training leads to models with perceptually aligned gradients. In this work, we explore the opposite question – *Do perceptually aligned gradients imply robustness?*

3. Do PAG Imply Robustness?

As mentioned in Section 2.3, 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 PAG will improve its robustness.

To this end, we propose encouraging the input-gradients of a classifier f_θ to uphold PAG. Due to the nature of our research question, we need to disentangle PAG from robust training and verify whether the former implies the latter. It 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:

$$\begin{aligned}\mathcal{L}_{total}(\mathbf{x}, y) =& \mathcal{L}_{CE}\left(f_\theta(\mathbf{x}), y\right) + \\ & \lambda \sum_{y_t=1}^{C} \mathcal{L}_{cos}\left(\nabla_{\mathbf{x}} f_\theta(\mathbf{x})_{y_t}, g(\mathbf{x}, y_t)\right),\end{aligned} \tag{4}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss defined in Equation (2), 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:

$$\mathcal{L}_{cos}(\mathbf{v}, \mathbf{u}) = 1 - \frac{\mathbf{v}^\top \mathbf{u}}{\texttt{max}(\|\mathbf{v}\|_2 \cdot \|\mathbf{u}\|_2, \varepsilon)}, \tag{5}$$

where ε is a small positive value so as to avoid division by zero. We emphasize that, in contrast to robust training methods such as AT and randomized smoothing (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, 2018b; 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 of any sort, making it suitable for exploring our titular research question.

After training a model to minimize the objective in Equation (4), 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 large-ϵ targeted PGD on test set images and qualitatively assess the validity of the resulting visual modifications, as in (Tsipras et al., 2019). Afterward, we test the adversarial robustness of the said model and compare it with vanilla baselines. If it demonstrates more aligned gradients and, in return, favorable robustness accuracy, we will have promoted an affirmative answer to the 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, $g(\cdot, \cdot)$, an assumption that does not hold in practice, as there is no clear way of obtaining point-wise realizations of them. In the following section, we begin by presenting practical and straightforward heuristic methods for obtaining approximations for these gradients, which we then use for training PAG-promoting classifiers. Next, we utilize diffusion models for obtaining theoretically justified such gradients as a better source of perceptually aligned gradients.

4. How are "Ground Truth" PAG Obtained?

In order to train a classifier for minimizing the objective in Equation (4), a "ground truth" perceptually aligned gradi-

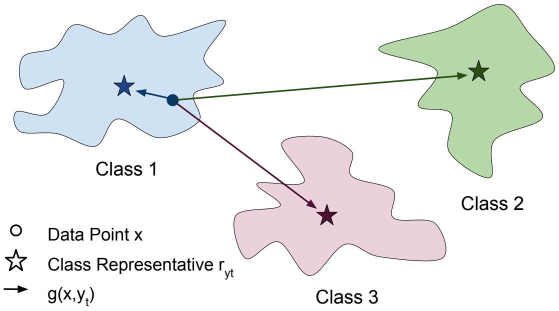


Figure 3. **Target class representatives**. An illustration of the heuristic realization of perceptually meaningful gradients.

ent $g(\mathbf{x}, y_t)$ needs to be provided for each training image $\mathbf{x} \in D$ and for each target class $y_t \in \{1, 2, \ldots, C\}$. Since a true such gradient is challenging to obtain, we instead explore a few general pragmatic approaches for approximating these PAGs, beginning with heuristic approaches and then advancing to theoretically justified ones.

4.1. 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 3, can be manifested in various ways, of which we consider the following:

One Image (OI): Set $\mathbf{r}_{y_t}$ to be an arbitrary training set image with label y_t, and use it as a global destination of y_t-targeted gradients.

Class Mean (CM): 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.

Nearest Neighbor (NN): 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

$$\mathbf{r}(\mathbf{x}, y_t) = \underset{\hat{\mathbf{x}} \in D_{y_t} \text{ s.t. } \hat{\mathbf{x}} \neq \mathbf{x}}{\arg\min} \|\hat{\mathbf{x}} - \mathbf{x}\|_2, \tag{6}$$

where D_{y_t} is the set of sample images with class y_t.

4.2. Score-Based Gradients

In this section, we describe our principled approach for approximating perceptually aligned gradients via Denoising Diffusion Probabilistic Models (DDPMs), which recently emerged as an interesting generative technique (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020). These models define noisy versions of an image $\mathbf{x}$, noted as $\{\mathbf{x}_t\}_{t=1}^{T}$, and their corresponding noisy data distributions $\{p_t(\mathbf{x}_t)\}_{t=1}^{T}$.[1] Sampling is performed through an iterative process that starts from a Gaussian noise and follows the direction of the *score function*, defined as $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$ and estimated by a neural network. Other works (Ho et al., 2022; Dhariwal & Nichol, 2021) have proposed to provide the class information to such networks, enabling them to model a class-conditional score function $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|y)$. We observe the similarity between the class-conditional score function and classification loss gradients w.r.t. the input image and hypothesize that gradients distilled from DDPM can be an improved source for perceptually aligned gradients ($g(\cdot, \cdot)$ in Equation (4)). We factorize the class-conditional score function via Bayes' rule as follows:

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|y) = \nabla_{\mathbf{x}_t} \log p(y|\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t), \tag{7}$$

leading to

$$\nabla_{\mathbf{x}_t} \log p(y|\mathbf{x}_t) = \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|y) - \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t). \tag{8}$$

This equation brings forth a novel usage of diffusion models – a principled way to estimate the correct gradients for the expression $\log p(y|\mathbf{x}_t)$. However, classification networks operate on clean images ($\mathbf{x}$) rather than noisy ones ($\mathbf{x}_t$). Therefore, in order to connect classifier input-gradients to DDPMs, we assume that $\log p(y|\mathbf{x}) \approx \log p(y|\mathbf{x}_t)$, for specific noise levels t. As a result, the desired estimation for "ground-truth" classifier input-gradients can be obtained by subtracting an unconditional score function from a class-condition one. The choice of t when distilling gradients via this approach introduces a tradeoff – too large values lead to gradients irrelevant to the input image, while too small ones lead to perceptually meaningless ones (in low noise levels, the conditional and unconditional scores are almost identical). Thus, we set t to be of medium values, yielding both perceptually and image-relevant gradients. We refer to this technique as Score-Based Gradients (SBG). We provide additional background and implementation details in Appendices B and G.2, respectively.

5. Experimental Results

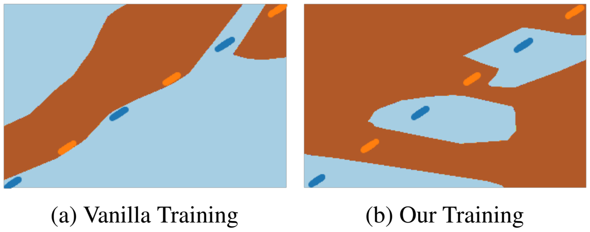
In this section, we empirically assess whether promoting PAG in classifiers improves the adversarial robustness at test time. We experiment using both synthetic and real datasets

[1] The dependence of $p_t(\cdot)$ on t is omitted to simplify notations.

and present our findings in the following section. Moreover, we study the correlation between PAG and robustness, *i.e.*, whether more aligned gradients yield improved robustness.

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* method with the vanilla training scheme that minimizes the cross-entropy loss. We train a two-layer fully-connected classifier twice: with our nearest neighbor method and 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 4. The baseline training method results in Figure 4a yields dimpled manifolds as decision boundaries, as hypothesized by (Shamir et al., 2021) – the decision boundary of DNN is very close to the data manifold, exposing the model to malicious perturbations. In contrast, in Figure 4b, 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%. 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. Therefore, our empirical findings strongly attest that PAG imply robustness in the synthetic use case. We provide additional experimental details in Appendix G.1.


(a) Vanilla Training

(b) Our Training

Figure 4. **Decision boundary visualization on the toy dataset**. 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 4a and 4b present the decision boundary of a vanilla training method and ours, respectively.

5.2. Real Datasets

With the encouraging findings presented in Section 5.1, we now turn to conduct thorough experiments to verify if indeed promoting PAG leads to improved adversarial robustness on real datasets. To methodologically answer this question, we train models to minimize Equation (4) and follow a two-step evaluation procedure. First, we validate that models trained with our approach obtain PAG, and next, we test their robustness. As for examining if a model has PAG, we qualitatively probe whether modifying an image to maximize a certain class probability, estimated by a model, leads to a meaningful class-related change. For evaluating the robustness, we adopt AutoAttack (AA) (Croce & Hein, 2020a) using both the L_∞ and L_2 threat models. Specifically, we compare models trained to minimize our objective, using the different realizations of PAG proposed in Section 4, with standard (vanilla) classifiers. Moreover, as a reference point, we also compare the results with adversarially trained (AT) (Madry et al., 2018) models. To establish empirical findings beyond a specific dataset and architecture, we experiment with models from different architecture families – Convolutional Neural Networks and Vision Transformers (ViT) (Dosovitskiy et al., 2021), and multiple datasets – CIFAR-10, STL, and CIFAR-100. In addition, we conduct an extensive architecture ablation in Appendix C.

CIFAR-10: First, we show in Figure 1 and Appendix I.1 that while vanilla models do not exhibit semantically meaningful changes, our approach does, as intended. Surprisingly, although our method is trained to have aligned gradients to some ground truth ones only on the data points, the model generalizes to have meaningful ones beyond these points. We proceed by quantitatively evaluating the performance on clean and adversarial versions of the test set and report the results in Table 1. As for the robustness evaluation, we utilize AutoAttack using $\epsilon = 8/255$ for L_∞ and $\epsilon = 0.5$ for L_2. While the vanilla baseline is utterly vulnerable to adversarial examples, all the tested PAG-inducing techniques improve the adversarial robustness substantially while maintaining competitive clean accuracies. It strongly suggests that promoting PAG can improve the classifier's robustness in real image datasets. A closer inspection of the results indicates that 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. Moreover, while both the heuristic-based and the theoretical-based "ground-truth" gradient sources substantially increase the robustness, the latter leads to significantly improved performance. Hence, in the following experiments, we mainly focus on the SBG approach.

Interestingly, the robustification obtained by Score-Based Gradients (SBG) is comparable to AT, without training on adversarial perturbations, potentially setting the foundations for non-adversarial methods for robust training. As our method does not perform adversarial training, it is faster than AT by up to ×6.14 (see Appendix H). In addition, SBG outperforms AT L_2 on unseen attacks (L_∞) on both RN-18 and ViT without training on adversarial perturbed images. Additional fascinating results are obtained on ViT – despite their tremendous popularity, ViTs are understudied in the

Table 1. CIFAR-10 results using ResNet-18 and ViT.

Arch.	Method	Clean	AA L_2 $\epsilon = 0.5$	AA L_∞ $\epsilon = 8/255$
RN-18	Vanilla	**93.61%**	00.00%	00.00%
	OI	79.46%	46.63%	13.50%
	CM	81.41%	47.25%	11.24%
	NN	80.65%	42.12%	07.51%
	SBG	78.56%	55.39%	$\underline{23.97\%}$
	AT L_∞	82.49%	$\underline{56.57\%}$	**37.59%**
	AT L_2	$\underline{86.79\%}$	**60.82%**	19.63%
ViT	Vanilla	$\underline{80.51\%}$	00.87%	00.01%
	OI	78.06%	15.47%	00.21%
	CM	78.98%	13.73%	00.17%
	NN	79.00%	13.91%	00.15%
	SBG	**81.28%**	**57.80%**	$\underline{22.85\%}$
	AT L_∞	62.20%	42.80%	**24.62%**
	AT L_2	72.81%	$\underline{42.99\%}$	08.13%

Table 2. Results on additional datasets using ResNet-18.

Dataset	Method	Clean	AA L_2	AA L_∞
STL	Vanilla	**82.60%**	00.00%	00.00%
	CM	70.66%	$\underline{58.90\%}$	$\underline{33.71\%}$
	SBG	$\underline{74.79\%}$	**65.96%**	**43.53%**
	AT L_∞	54.90%	46.33%	28.30%
	AT L_2	54.99%	46.04%	23.33%
CIFAR-100	Vanilla	**74.36%**	00.00%	00.00%
	CM	$\underline{58.89\%}$	19.94%	02.78%
	SBG	55.94%	$\underline{29.25\%}$	$\underline{08.24\%}$
	AT L_∞	52.92%	26.31%	**14.63%**
	AT L_2	58.05%	**30.51%**	08.03%

field of adversarial robustness, which still mainly focuses on CNNs. Similar to the findings of (Mo et al., 2022), our results show that AT is significantly less effective for training ViTs from scratch. On the contrary, SBG outperforms it both in robustness and clean accuracy. Specifically, it surpasses the clean accuracy of AT L_2 by $+8.47\%$ and the robust accuracy to L_2 attacks by $+14.81\%$. Besides the improved quantitative performance, Figure 10 in Appendix I shows that AT leads to inferior PAG compared to our method, which further attests to the strong bidirectional connection between PAG and adversarial robustness. In addition, the clean accuracy obtained by SBG is higher than the “vanilla” one (81.28% compared to 80.51%). We hypothesize that the improved performance of SBG in the ViT case stems from the fact that it serves as a beneficial regularization. Such a regularization is required as CIFAR-10 is a relatively small dataset, and ViT is highly expressive.

STL: To better validate that the connection between PAG and robustness holds in general, we test our approach on STL (Coates et al., 2011), which contains images of a higher resolution of 96×96 pixels. Besides its resolution, we choose STL mainly due to its relatively small size – $5,000$ training and $8,000$ test images. While it is known that low data regimes are Adversarial Training’s Achilles’ heel, as it requires more training data (Schmidt et al., 2018; Zhai et al., 2019), we aim to investigate the connection between PAG and robustness in such a challenging setup. We conduct similar experiments with ResNet-18 as in CIFAR-10 and consider $\epsilon = 4/255$ for L_∞ and $\epsilon = 0.5$ for L_2 as the threat models. We summarize our qualitative PAG results in Figure 6 in Appendix I and our quantitative ones in Table 2. As can be seen in the results, both the heuristic and SBG approaches yield models substantially more robust than standard training. Moreover, interestingly, while AT struggles to obtain decent results, both our heuristic and principled approaches significantly outperform it in clean and adversarial accuracy, while SBG is substantially better. Specifically, SBG outperforms AT in clean accuracy by a staggering $+19.8\%$ and in L_2 robustness by $+19.63\%$.

CIFAR-100: While both CIFAR-10 and STL contain only 10 classes, we aim to examine whether our approach is also effective on datasets with more classes. To this end, we experiment with ResNet-18 on the CIFAR-100 dataset (Krizhevsky et al.), containing 100 classes. To reduce the computational cost, we minimize our objective in 4 using a subset of 10 classes, randomly selected in each training iteration. As for the robustness settings, we use the same threat models as in CIFAR-10. As can be seen in Table 2 and similar to the trends in CIFAR-10 and STL, classifiers with PAG are significantly more robust than “vanilla” ones, attesting to the bidirectional connection between PAG and robustness. Moreover, SBG obtains similar performance as AT L_2 without performing adversarial training.

5.3. Robustness To Large ϵ Attacks

Despite the great success of adversarial training (AT) in obtaining robust models, one of its well-known limitations is their poor generalization to unseen attacks. Specifically, as AT is performed using a specific threat model, it performs satisfactorily against attacks from the trained threat model but can be easily circumvented by different types of attacks (*e.g.*, larger ϵ). This limitation was listed by (Hendrycks et al., 2021; Bai et al., 2021) as one of the major unsolved problems of adversarial training methods. As our approach does not perform training on adversarial examples, we hypothesize that it can cope better against unseen attacks. To examine this, we first train a ResNet-18 on CIFAR-10 using our SBG approach, standard (“vanilla”), and adversarial training using both L_2 and L_∞ with ϵ of 0.5 and $8/255$, respectively. Next, we examine their performance against

Table 3. Large attacks on the CIFAR-10 dataset using ResNet-18.

Method	Clean	AA L_2 $\epsilon = 1$	AA L_∞ $\epsilon = 16/255$
Vanilla	**93.61%**	00.00%	00.00%
SBG	<u>78.56%</u>	**30.43%**	<u>01.92%</u>
AT L_∞	82.49%	<u>25.26%</u>	**08.02%**
AT L_2	<u>86.79%</u>	18.81%	00.50%

Table 4. Improving AT via gradient alignment.

Method	Clean	AA L_2 $\epsilon = 0.5$	AA L_∞ $\epsilon = 8/255$
AT L_2	86.79%	60.82%	19.63%
+ ours	85.54%	61.73%	23.52%
Δ	**-1.25**	**+0.91**	**+3.89**
TRADES L_∞	83.04%	57.41%	43.54%
+ours	84.59%	62.66%	45.78%
Δ	**+1.55**	**+5.25**	**+2.24**

two types of unseen attacks – larger ϵ and different threat models (different norms) and summarize our results in Table 3. As can be seen, our approach is substantially more robust to unseen attacks. Specifically, although SBG obtains 55.39% on $\epsilon = 0.5$, compared to 60.82% of AT L_2, it significantly outperforms it under L_2 attacks using $\epsilon = 1.0$. Of course, training AT L_2 against $\epsilon = 1.0$ would result in a more robust model against this threat model, but at a high cost of clean accuracy. Contrary, our method does not train against any attack and shows an impressive generalization to different threat models. It is surprising that our method generalizes well far beyond the data samples, as it trains on the data points themselves rather than on ϵ volume as AT. These results show an additional and intriguing benefit of our proposed approach – robustness to unseen attack.

5.4. On The Correlation Between PAG and Robustness

In previous sections, we discover that models that possess PAG are inherently more robust than ones that do not. Our framework enables us to examine whether there is a correlation between the level of the classifier's gradient alignment and its robustness. In particular, do models with more aligned gradients more robust? To empirically answer this question, we train classifiers using SBG with different values of λ (*i.e.*, PAG regularization hyperparameter) and conclude our results in Figure 5. As can be seen, increasing λ yields models with more aligned gradients, as the modified images are more semantically similar to the different target classes. Interestingly, such models are also significantly more robust than the ones with less aligned gradients. These findings attest that PAG and robustness are highly correlated, and models with more PAG are substantially more robust.

6. Improved AT Via Gradient Alignment

The results in Section 5.4 show that increasing the level of PAG results in more robust models. As adversarial training methods are known to yield models that possess PAG, it brings the question of whether such methods can be further improved by increasing the level of gradient alignment. Specifically, this can be explored by introducing our PAG-inducing objective into adversarial training techniques. In particular, we experiment on CIFAR-10 with AT (Madry et al., 2018) using $L_2, \epsilon = 0.5$ and TRADES (Zhang et al., 2019) using $L_\infty, \epsilon = 8/255$, with ResNet-18 and Wide ResNet 34-10, respectively. Our results (Table 4) show that introducing our loss to adversarial training techniques increase their robustness against the trained threat model by $+0.91\%$, $+2.24\%$, for AT L_2 and TRADES L_∞, respectively. Moreover, it increases, even more, their robustness against the unseen attack by $+3.89\%$, $+5.25\%$, for AT L_2 and TRADES L_∞, respectively. It bodes well with the generalization findings in Section 5.3. We provide additional details in Appendix G.3 and robustness evaluation against additional attacks in Table 6.

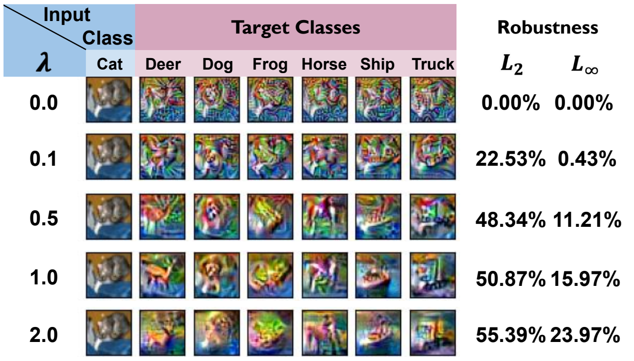


Figure 5. **Correlation between PAG and robustness**. Demonstration of the positive correlation between PAG and robustness. Higher λ values lead to more PAG and in return, more robustness.

7. Conclusions and Future Work

While previous work demonstrates that adversarially robust models uphold the PAG property, in this work, we 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 strongly suggest that promoting PAG leads to more robust models, exposing the bidirectional connection between the two. Interestingly, our

SBG approach achieves comparable robustness to AT and even outperforms it in the following setups – (i) low data regimes; (ii) ViTs; (iii) large ϵ attacks. In addition, we discover the positive correlation between PAG and robustness and harness this insight to improve existing robustification techniques substantially. We believe that the obtained results can be further improved with better realizations of PAG and a more sophisticated method, and we hope this work will serve as a foundation for such a line of research.

8. Acknowledgements

This research was partially supported by the Israel Science Foundation (ISF) under Grant 335/18, the Crown Family Foundation Doctoral Fellowship, and the Council for Higher Education - Planning & Budgeting Committee.

References

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

Amit, T., Nachmani, E., Shaharabany, T., and Wolf, L. Segdiff: Image segmentation with diffusion probabilistic models. *CoRR*, 2021.

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

Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. Synthesizing robust adversarial examples. In *International Conference on Machine Learning, ICML*, 2018.

Avrahami, O., Lischinski, D., and Fried, O. Blended diffusion for text-driven editing of natural images. In *Conference on Computer Vision and Pattern Recognition, CVPR*, 2022.

Bai, T., Luo, J., Zhao, J., Wen, B., and Wang, Q. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021.

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*, 2013.

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

Carlini, N. and Wagner, D. A. Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM Workshop on Artificial Intelligence and Security, AISec@CCS*, 2017b.

Chan, A., Tay, Y., and Ong, Y. What it thinks is important is important: Robustness transfers through input gradients. In *2020 Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020.

Coates, A., Ng, A., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics, AIStat*, 2011.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning, ICML*, 2019a.

Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning, ICML*, 2019b.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning, ICML*, 2020a.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning, ICML*, 2020b.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition, CVPR*, 2009.

Dhariwal, P. and Nichol, A. Q. Diffusion models beat GANs on image synthesis. In *Neural Information Processing Systems, NeurIPS*, 2021.

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 *Conference on Computer Vision and Pattern Recognition, CVPR*, 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. In *International Conference on Learning Representations, ICLR*, 2021.

Efron, B. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 2011.

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., Santurkar, S., Tsipras, D., Tran, B., and Madry, A. Adversarial robustness as a prior for learned representations. *arXiv*, 2019.

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, ICML*, 2019.

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

Ganz, R. and Elad, M. Bigroc: Boosting image generation via a robust classifier. *CoRR*, 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. In *Conference on Computer Vision and Pattern Recognition, CVPR*, 2016.

Hendrycks, D., Carlini, N., Schulman, J., and Steinhardt, J. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Neural Information Processing Systems, NeurIPS*, 2020.

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., and Salimans, T. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research, JMLR*, 2022.

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.), *Neural Information Processing Systems, NeurIPS*, 2020.

Jakubovitz, D. and Giryes, R. Improving dnn robustness to adversarial attacks using jacobian regularization. In *European Conference on Computer Vision, ECCV*, 2018.

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

Kawar, B., Vaksman, G., and Elad, M. SNIPS: Solving noisy inverse problems stochastically. *Neural Information Processing Systems, NeurIPS*, 2021.

Kawar, B., Elad, M., Ermon, S., and Song, J. Denoising diffusion restoration models. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022a.

Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., and Irani, M. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022b.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-100 (canadian institute for advanced research).

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems, NeurIPS*, 2012.

Kurakin, A., Goodfellow, I. J., and Bengio, S. Adversarial examples in the physical world. In *International Conference on Learning Representations, ICLR 2017, Workshop Track Proceedings*, 2017.

Le, Y. and Yang, X. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Lécuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with differential privacy. In *Symposium on Security and Privacy, SP*, 2019.

Li, B., Chen, C., Wang, W., and Carin, L. Certified adversarial robustness with additive noise. In *Neural Information Processing Systems, NeurIPS*, 2019.

Liu, X., Park, D. H., Azadi, S., Zhang, G., Chopikyan, A., Hu, Y., Shi, H., Rohrbach, A., and Darrell, T. More control for free! image synthesis with semantic diffusion guidance. *CoRR*, 2021.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations, ICLR*, 2018.

Miyasawa, K. An empirical Bayes estimator of the mean of a normal population. *Bull. Inst. Internat. Statist.*, 1961.

Mo, Y., Wu, D., Wang, Y., Guo, Y., and Wang, Y. When adversarial training meets vision transformers: Recipes from training to architecture. *CoRR*, abs/2210.07540, 2022. doi: 10.48550/arXiv.2210.07540. URL https://doi.org/10.48550/arXiv.2210.07540.

Nguyen, A. M., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Conference on Computer Vision and Pattern Recognition, CVPR*, 2015.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In Meila, M. and Zhang, T. (eds.), *International Conference on Machine Learning, ICML*, 2021.

Pang, T., Yang, X., Dong, Y., Xu, T., Zhu, J., and Su, H. Boosting adversarial training with hypersphere embedding. In *Neural Information Processing Systems, NeurIPS*, 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 *Neural Information Processing Systems, NeurIPS*, 2019.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

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

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

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

Ruderman, D. L. The statistics of natural images. *Network: Computation in Neural Systems*, 1994.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. *CoRR*, 2022.

Salman, H., Li, J., Razenshteyn, I. P., Zhang, P., Zhang, H., Bubeck, S., and Yang, G. Provably robust deep learning via adversarially trained smoothed classifiers. In *Neural Information Processing Systems, NeurIPS*, 2019.

Santurkar, S., Tsipras, D., Tran, B., Ilyas, A., Engstrom, L., and Madry, A. *Image Synthesis with a Single (Robust) Classifier*. 2019.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In *Neural Information Processing Systems, NeurIPS*. Curran Associates, Inc., 2018.

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

Shao, R., Yi, J., Chen, P., and Hsieh, C. How and when adversarial robustness transfers in knowledge distillation? *CoRR*, abs/2110.12072, 2021.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning, ICML*, 2015.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. In *Neural Information Processing Systems, NeurIPS*, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations, ICLR*, 2021.

Stein, C. M. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, 1981.

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]. *Signal Processing Magazine*, 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. In *International Conference on Machine Learning and Applications, ICMLA*, 2018.

Theis, L., Salimans, T., Hoffman, M. D., and Mentzer, F. Lossy compression with gaussian diffusion. *CoRR*, 2022.

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021.

Tramèr, F., Carlini, N., Brendel, W., and Madry, A. On adaptive attacks to adversarial example defenses. In *Neural Information Processing Systems, NeurIPS*, 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.

Vahdat, A., Kreis, K., and Kautz, J. Score-based generative modeling in latent space. In *Neural Information Processing Systems, NeurIPS*, 2021.

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., and Gu, Q. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations, ICLR*, 2020.

Xie, C., Wu, Y., van der Maaten, L., Yuille, A. L., and He, K. Feature denoising for improving adversarial robustness. In *Conference on Computer Vision and Pattern Recognition, CVPR*, 2019.

Zhai, R., Cai, T., He, D., Dan, C., He, K., Hopcroft, J. E., and Wang, L. Adversarially robust generalization just requires more unlabeled data. *CoRR*, 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 *International Conference on Machine Learning, ICML*, 2019.

A. Related Work

Input-Gradient Penalty Approaches 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). This line of work attests that requiring small gradient norms, regardless of their direction, leads to robustness. Moreover, none of these works promotes nor exhibits perceptually aligned gradients. In opposition to the above methods, we penalize the direction of the gradients, regardless of their norm. By showing that promoting PAG improves robustness, our method can be viewed as an alternative input-gradient loss for improving robustness. Nevertheless, the main goal of our work is to better study PAG and its connection with adversarial robustness.

Robust Knowledge Distillation Approaches The robust input-gradient distillation methods (Chan et al., 2020; Shao et al., 2021) is an additional related line of work. 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 unsuitable for assessing whether perceptually-aligned gradients inherently promote robustness, as they implicitly rely on prior adversarial training.

B. Denoising Diffusion Probabilistic Models

Denoising diffusion probabilistic models (DDPMs) are a new fascinating generative approach (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020). Such methods achieve state-of-the-art results in image generation (Dhariwal & Nichol, 2021; Song et al., 2021; Vahdat et al., 2021), and were additionally deployed in several downstream tasks such as inverse problems (Kawar et al., 2021; 2022a), image compression (Theis et al., 2022), image segmentation (Amit et al., 2021), image editing (Liu et al., 2021; Avrahami et al., 2022; Kawar et al., 2022b), and text-to-image generation (Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022), among others.

The core idea of these models, also known as score-based generative models, is to start from a random Gaussian noise image $\mathbf{x}_T$ and then iteratively denoise it into a photorealistic image $\mathbf{x}_0$ in a controlled manner. This process can also be interpreted as an annealed version of Langevin dynamics (Song & Ermon, 2019), where each iteration $t \in \{T, T-1, \ldots, 1\}$ follows the direction of the *score function*, defined as $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$, with an additional noise for stochasticity. Each intermediate image $\mathbf{x}_t$ can be considered a noisy version of a pristine image $\mathbf{x}_0$, with a pre-defined noise level σ_t. The score function can be estimated using a neural network trained for mean-squared-error denoising (Stein, 1981; Miyasawa, 1961; Efron, 2011). This estimation can also be generalized for denoising models conditioned on a class label y, obtaining $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|y)$ (Ho et al., 2022). In this work, we propose a novel usage of diffusion models for approximating perceptually aligned gradients and train classifiers to be aligned with them. Our experiments show that this principled PAG realization is highly reliable, leading to a substantially increased robustness.

ViT: As can be seen in Figure 10, while "vanilla" classifiers do not possess PAG, both our heuristic approaches and AT models obtain more aligned gradients. Interestingly, our principled SBG approach leads to the best PAG among the tested methods and obtains the highest robustness. This further attests to the positive correlation between PAG and robustness.

B.1. STL

In this section, we qualitatively examine the existence of PAG for ResNet-18 on the STL dataset. Specifically, we consider standard-trained classifiers ("vanilla), adversarially trained ones (AT), and models trained using both our heuristic (OI, CM, and NN) and principled (SBG) approaches. As can be seen in Figure 6, while the "vanilla" classifier does not possess PAG at all and AT models feature a minimal gradient alignment, our approaches lead to much more aligned gradients. As in CIFAR-10, we visualize additional demonstrations for the SBG case in Figure 12. This further strengthens the connection between PAG and robustness, as our methods achieve much higher robustness than adversarial training on this dataset.

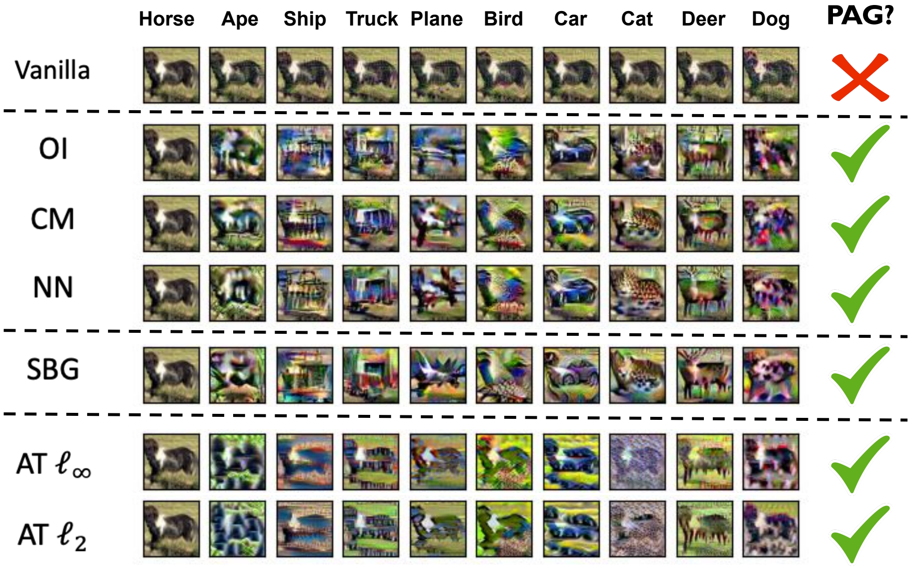


Figure 6. **ResNet-18 (RN-18) PAG examination on STL**. PAG examination for ViT on CIFAR-10 trained standardly ("vanilla"), adversarially (AT) and using both our heuristic (OI, CM and NN) and principled approaches (SBG).

C. Additional Architectures Ablation

In this section, we provide the results of applying our method to additional architecture types. While we focus in Section 5 on skip-connection-based convolutional NN (ResNet-18) and an attention-based one (ViT), we turn to examine it on other types of architectures. Specifically, we apply it to VGG (Simonyan & Zisserman, 2014), a convolutional network without residual connection, and MLP Mixer (Tolstikhin et al., 2021), a top-performing dense architecture. Empirically proving that perceptually aligned gradients imply robustness in these architectures will further strengthen that the connection between the two is model agnostic. We experiment with such architectures using the SBG approach on the CIFAR-10 dataset and report the results in Table 5. We evaluate the performance on clean and adversarially perturbed images using AutoAttack with a $L_2, \epsilon = 0.5$ threat model. For the MLP Mixer, we follow the CIFAR-10 adjusted implementation[2] and train it for 100 epochs using a batch size of 128. As for our approach, we utilize the SBG PAG realizations and set $\lambda = 0.5$. As for the VGG, we follow the implementation for CIFAR-10[3] and train for 100 epochs using a batch size of 64. Similarly to MLP Mixer, we use SBG with $\lambda = 0.5$. Moreover, we experiment with VGG-11, VGG-13, and VGG-16 to further study the depth effect. As our results suggest, the connection between PAG and robustness is general and architecture-independent.

Table 5. Accuracy on CIFAR-10 using VGG and MLP Mixer architectures.

Method	Arch.	Clean	AutoAttack L_2
Vanilla	VGG-16	**92.32%**	00.20%
SBG		81.93%	**42.03%**
Vanilla	VGG-13	**92.47%**	00.11%
SBG		82.05%	**41.49%**
Vanilla	VGG-11	**90.82%**	02.50%
SBG		79.22%	**35.79%**
Vanilla	MLP-Mixer	**72.05%**	00.50%
SBG		63.04%	**35.97%**

[2] https://github.com/omihub777/MLP-Mixer-CIFAR
[3] https://github.com/chengyangfu/pytorch-vgg-cifar10

D. Robustness to Additional Adversarial Attacks

In the main paper, we test our approach on L_2 and L_∞ norm-bounded adversarial attacks and compare it to Adversarial Training. In this section, we extend the evaluation to additional adversarial attacks, including non-norm bounded ones, using Foolbox[4] and report the results in Table 6. Specifically, we use ResNet-18 trained on CIFAR-10 using SBG and Adversarial Training with L_2 threat model and $\epsilon = 0.5$.

Table 6. Robustness of ResNet-18 trained on CIFAR-10 against various adversarial attacks.

Attack	AT L_2	SBG
SaltAndPepperNoiseAttack	29.86%	**42.59%**
BinarySearchContrastReductionAttack	20.02%	**44.71%**
GaussianBlurAttack	28.03%	**51.54%**
LinearSearchBlendedUniformNoiseAttack	**34.95%**	34.30%
L0FMNAttack	**43.03%**	40.96%
L1FMNAttack	31.60%	**40.24%**
EADAttack	8.57%	**33.64%**

E. Tiny ImageNet Experiments

To further show that promoting PAG increases adversarial robustness, we conduct experiments on the challenging Tiny ImageNet dataset (Le & Yang, 2015), containing $100,000$ $64 \times 64 \times 3$ images of 200 classes. We compare the obtained results of our CM and SBG approaches to the standard training and report the results in Table 7. Our results show that models with PAG are inherently more robust.

F. Visualization of PAG Realizations

We visualize our proposed realizations for approximating Perceptually Aligned Gradients in Figure 7 on CIFAR-10. In the top three rows, we show the results of the heuristic-based methods. A ghosting effect can be seen as these gradients derive from the subtraction of two images. However, in Score-Based Gradients, the modifications focus on the object, and features of the target class can be observed (i.e., horse features). The nature of SBG is object-centric as the performed modifications focus on the object itself, similar to AT.

G. Implementation Details

G.1. Toy Dataset

Data: We experiment with our approach on a 2-dimensional synthetic dataset to demonstrate its effects. To this end, we construct a dataset of 6,000 samples from two classes containing precisely 3,000 examples. Our samples, $\mathbf{x} = [x_1, x_2]$, reside on the straight line $x_2 - 2x_1 = 0$ in the 2-dimensional space $\mathbb{R}^2$, where each class $y \in \{0, 1\}$ follows a Gaussian mixture distribution. Each class contains three modes, and each 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 $\{-30, 10, 50\}$. This way, the cardinal manifold assumption according to which high-dimensional images reside on a lower-dimensional manifold (Ruderman, 1994) holds. To evaluate performance, we generate a balanced test set from the same distribution

[4] https://github.com/bethgelab/foolbox

Table 7. Results on Tiny ImageNet using ResNet-18.

Method	Clean	AA L_2	AA L_∞
Vanilla	**61.19%**	02.37%	00.00%
CM	50.04%	19.21%	03.14%
SBG	57.40%	**25.21%**	**05.50%**

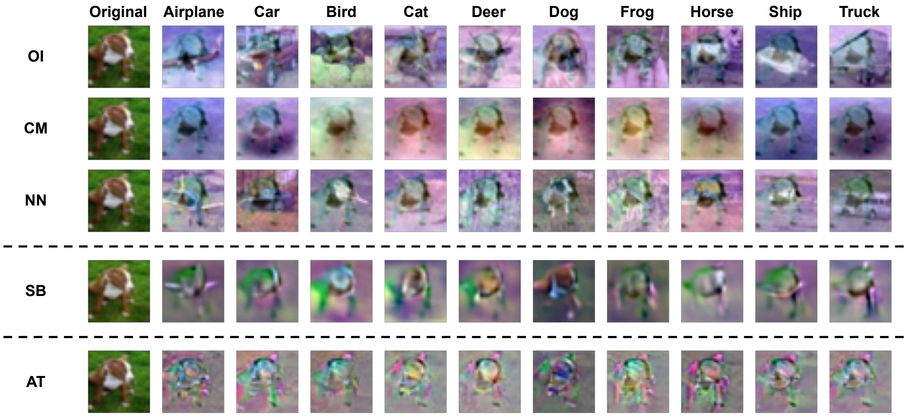


Figure 7. **Visualization of different realizations of PAG**. We visualize the input-gradients w.r.t. different target classes for the PAG realizations considered in this work – heuristic (OI, CM, NN) and principled (SBG) realizations. As a reference, we visualize the gradients of the adversarially trained classifier.

consisting of 600 samples.

Architecture and Training: We use a 2-layer fully-connected network $(2 \rightarrow 32 \rightarrow 2)$ 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 Tesla V100 GPU.

Evaluation: As detailed in the paper, we test the performance of the models using standard and adversarial evaluation. We draw 600 test samples from the same distribution as the train set for the standard one 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.

G.2. Real Datasets

Data: As for our real datasets experiments, we use CIFAR-10, CIFAR-100, and STL that contain images of size $32 \times 32 \times 3$ (the formers) and $96 \times 96 \times 3$ (the latter). For each dataset and PAG "ground-truth" realization, we construct a training set by computing C targeted gradients for each training sample ($C = 10$ for CIFAR-10 and STL and $C = 100$ for CIFAR-100) for reproducibility and consistency purposes.

To obtain our Score-Based Gradients (SBG), we follow the implementation of (Nichol & Dhariwal, 2021)[5] for training a class-conditional diffusion model for CIFAR-10, CIFAR-100, and STL datasets. We use their CIFAR-10 architecture for the CIFAR datasets and their ImageNet architecture for STL, adapting the image size by a simple bicubic interpolation. In particular, for a C-classes dataset, we train a single class-conditioned diffusion model with $C + 1$ classes, where the additional class represents the absence of class information and thus models the unconditional score function. Instances of this class are drawn with probability $1/C$, and they originate uniformly from each of the C classes. This way enables us to distill gradients using Equation (8) with the same model, mitigating the need to scale the outputs of the conditional and unconditional models.

As specified in Section 4.2, Equation (8) deals with noisy images, not clean ones, as we consider in this work. Nevertheless, picking a proper noise level t enables us to distill meaningful and valuable gradients. As explained in Section 4.2, too large values lead to gradients irrelevant to the input image, while too small ones lead to perceptually meaningless ones. However, setting t to an intermediate value leads to semantically meaningful outputs. We demonstrate this in Figure 8.

Training: For all the tested datasets, we train the classifier (ResNet-18 or ViT) 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

[5] https://github.com/openai/improved-diffusion

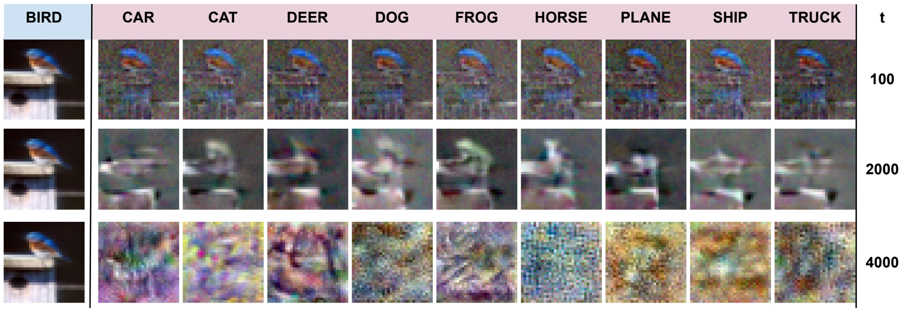


Figure 8. **Gradients realization using SBG with different t values**. We visualize the effects of different levels of t when distilling gradients using the SBG approach. As can be seen, while too large or small t values lead to meaningless gradients, mid-values of t leads to the desired outputs.

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 CIFAR-100 and 32 for STL. As for the ViT, we use a publicly available implementation, adjusted to CIFAR-10[6], containing 6.3 million parameters. We present in Table 8 the best choices of λ – the coefficient of our PAG promoting auxiliary loss term in all the tested datasets and methods. The values of λ suggest that higher values should be applied for better gradient sources (*e.g.*, SBG's ideal λ is higher than the heuristic methods)

As for our baselines, we use the same training hyperparameters mentioned above. Regarding the AT, we use 7 steps PGD using a step size of $1.5 * \frac{\epsilon}{7}$ and follow the base implementation presented in (Zhang et al., 2019)[7] and extend it to L_2.

Table 8. Values of the hyperparameter λ.

Method	CIFAR-10		STL	CIFAR-100
	RN-18	ViT	RN-18	RN-18
One Image	0.5	0.1	-	-
Class Mean	0.4	0.1	0.2	0.5
Nearest Neighbor	0.4	0.1	-	-
SBG	2	2	1	1

Computational Resources: We use two NVIDIA RTX A4000 16GB GPUs for each experiment.

Evaluation: We use the de-facto standard evaluation library of AutoAttack (Croce & Hein, 2020a)[8].

G.3. Improved AT Via Gradient Alignment

As demonstrated in Section 6, introducing gradient alignment regularization further improves the robustness achieved by robustification methods. In particular, we show this for Adversarial Training using $L2, \epsilon = 0.5$ and TRADES using $L_\infty, \epsilon = 8/255$, with ResNet-18 and Wide ResNet 34-10. We use the official code repository for the latter and implement our loss upon it. We train each model twice with the same hyperparameters – with and without our regularization. We set λ to be 0.2 for AT and 0.5 for TRADES. Next, we used AutoAttack to evaluate the performance of the trained models against both L_2 and L_∞ attacks.

H. Runtime Comparison with Adversarial Training

As our method does not compute adversarial examples (an iterative process), it is faster than adversarial training. To quantify this, we conduct a runtime comparison using a batch size of size 1, using ResNet-18 and the CIFAR-10 dataset, and reveal that our method is faster than AT-PGD-7 and AT-PGD-20 by x2.13 and x6.14, respectively. However, utilizing our approach with Score-Based-Gradients (SBG) requires access to a trained diffusion model, which takes additional time to

[6] https://github.com/omihub777/ViT-CIFAR
[7] https://github.com/yaodongyu/TRADES
[8] https://github.com/fra31/auto-attack

train and sample. Nevertheless, the same diffusion model is utilized to generate the SBG gradients used to train the different classifiers considered in this work. For example, training a diffusion model on the CIFAR-10 dataset to achieve satisfactory results takes several hours. Nevertheless, as we use a single t value for crafting the SBG, the diffusion training time can be significantly reduced by training only on the corresponding noise level.

I. Additional PAG Demonstrations

I.1. CIFAR-10

In this section, we provide a qualitative examination of the existence of PAG for ResNet-18 and ViT on the CIFAR-10 dataset. Specifically, we consider standard-trained classifiers ("vanilla), adversarially trained ones (AT), and models trained using both our heuristic (OI, CM, and NN) and principled (SBG) approaches.

ResNet-18: As can be seen in Figure 9, while "vanilla" classifiers do not possess PAG, both our approaches and AT models obtain aligned gradients. To better demonstrate the gradient obtained by SBG, we visualize in Figure 11 the output of applying strong targeted adversarial examples, starting from random images.

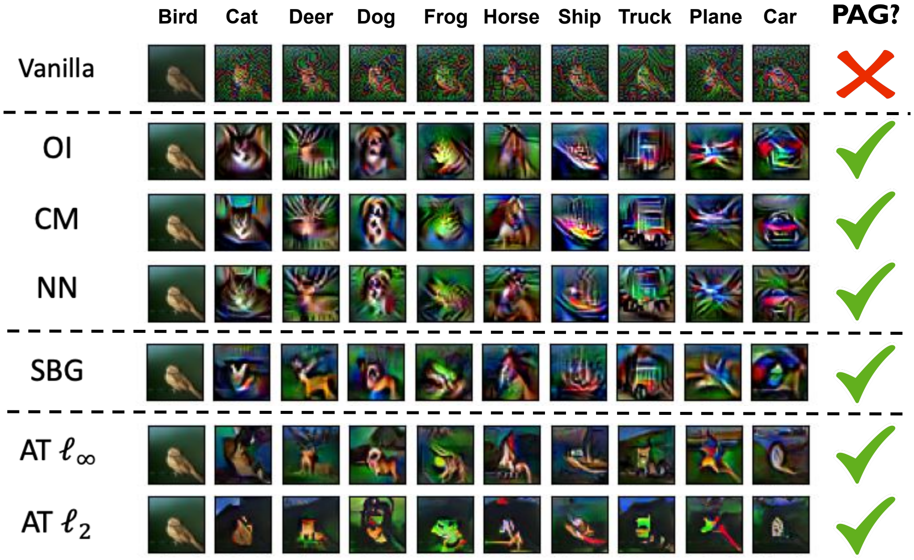


Figure 9. **ResNet-18 (RN-18) PAG examination on CIFAR-10**. Large-ϵ targeted attacks on RN-18 trained standardly ("vanilla"), adversarially (AT), and using both our heuristic (OI, CM, and NN) and principled approaches (SBG) on CIFAR-10.

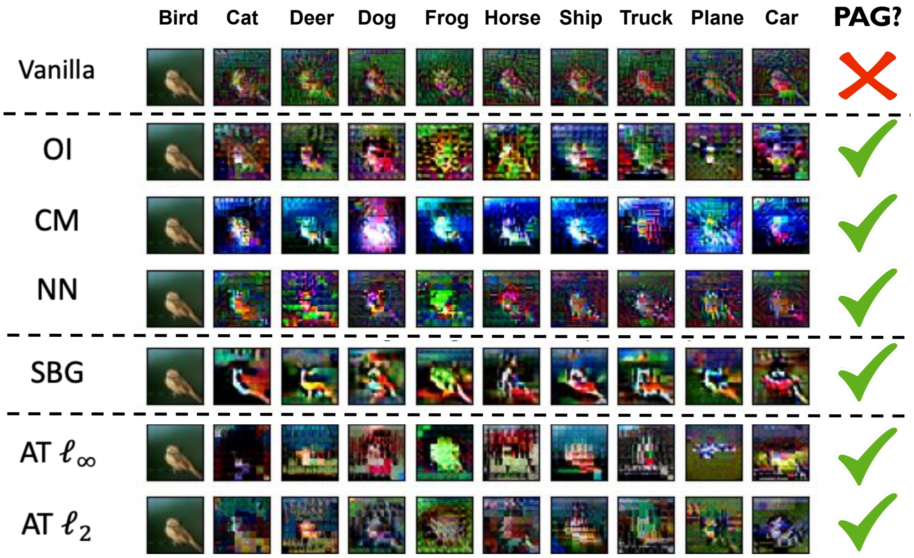


Figure 10. **Vision Transformer (ViT) PAG examination on CIFAR-10**. Large-ϵ targeted attacks on ViT trained using standardly ("vanilla"), adversarially (AT), and using both our heuristic (OI, CM, and NN) and principled approaches (SBG) on CIFAR-10.

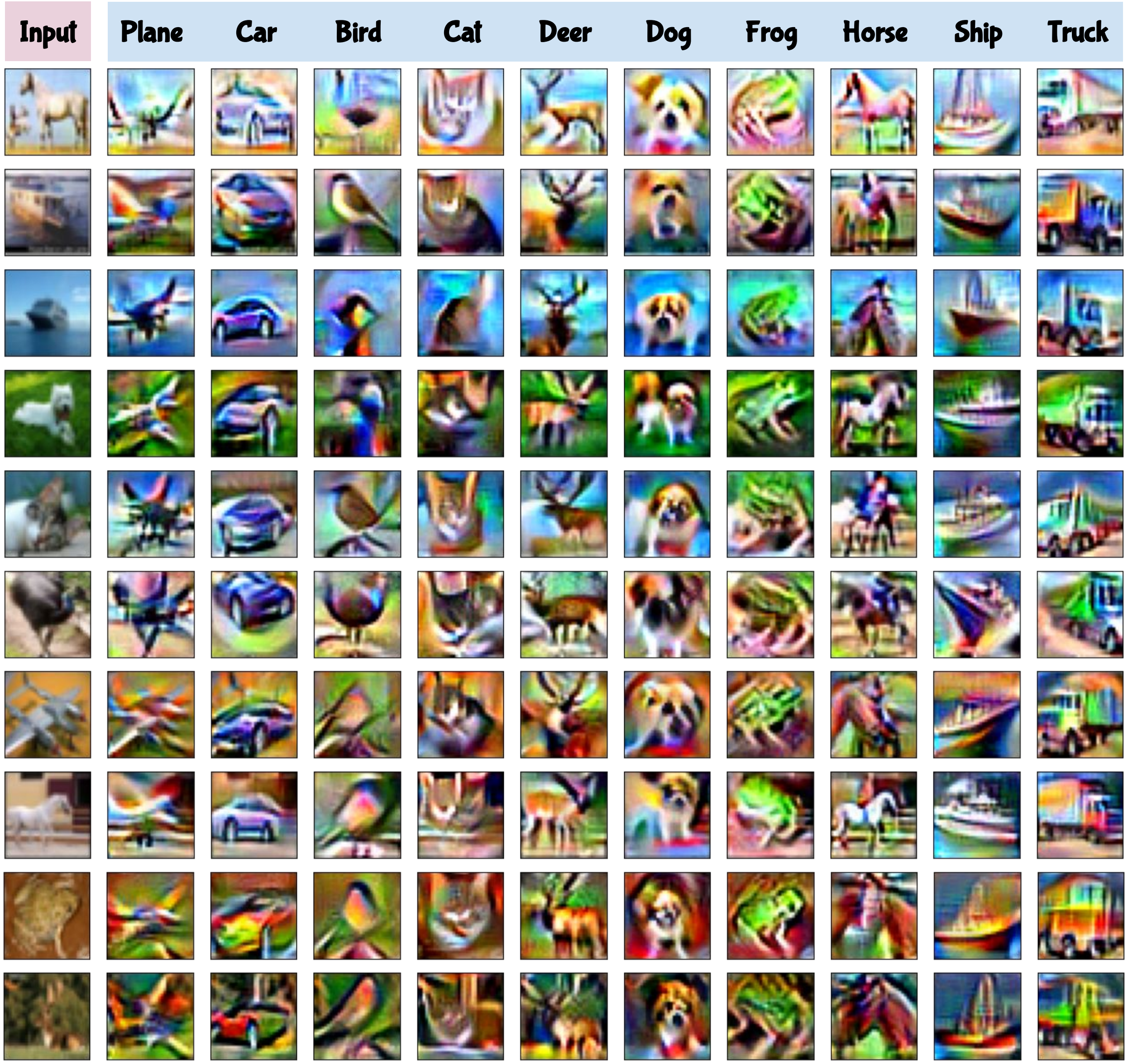


Figure 11. **CIFAR-10 PAG Visualizations of SBG**. Targeted large-ϵ adversarial examples on ResNet-18 trained on CIFAR-10 using SBG. We present 10 randomly selected images and transform them into all the possible classes. As can be seen, all images contain meaningful class-related semantic information.

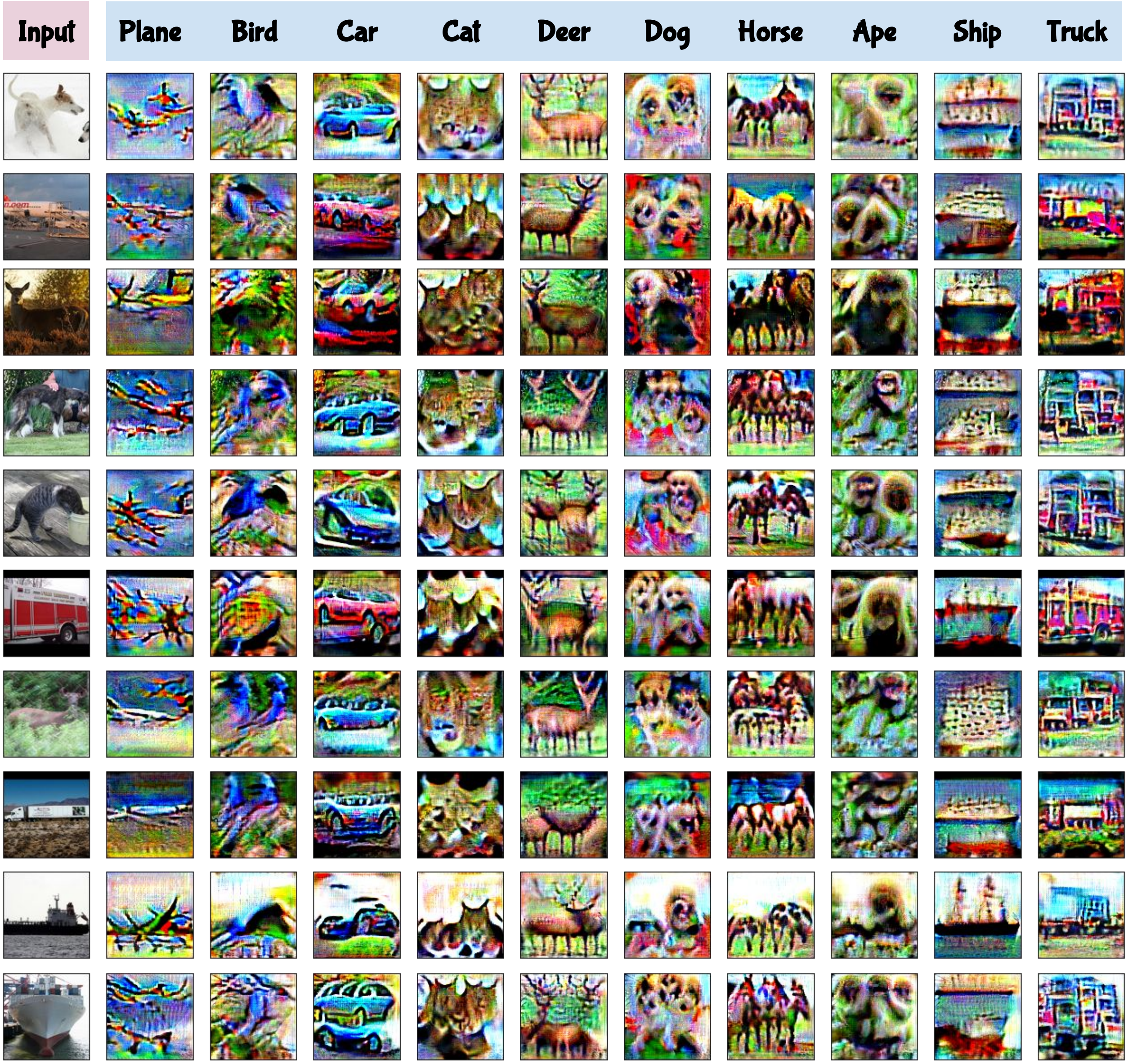


Figure 12. **STL PAG Visualizations of SBG**. Targeted large-ϵ adversarial examples on ResNet-18 trained on STL using SBG. We present 10 randomly selected images and transform them into all the possible classes. As can be seen, all images contain meaningful class-related semantic information.