Do Perceptually Aligned Gradients Imply Robustness?

Roy Ganz, Bahjat Kawar and Michael Elad

Technion, Israel

ICML 2023

TECHNION
Israel Institute of Technology

ICML
International Conference On Machine Learning

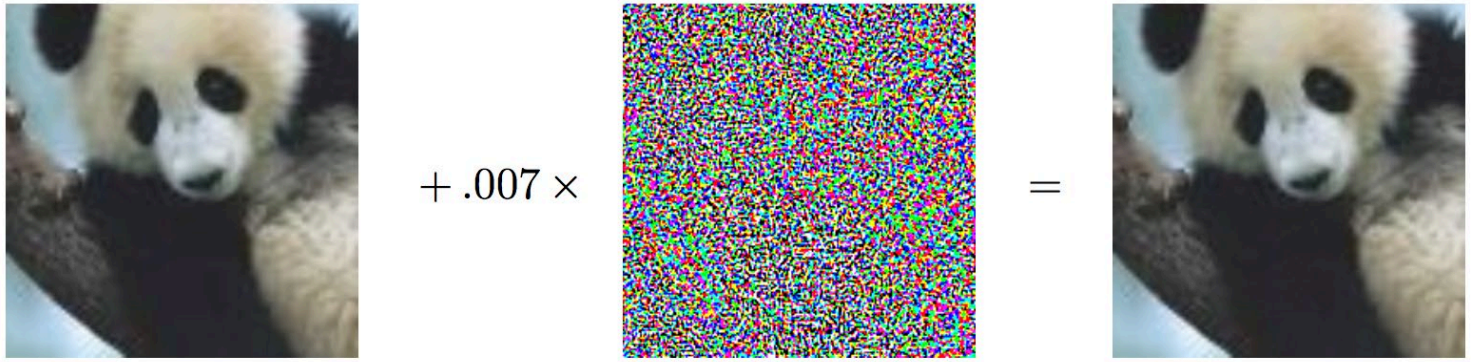
Background

Adversarial Robustness and Perceptually Aligned Gradients (PAG)

Background

Adversarial Robustness and Perceptually Aligned Gradients (PAG)

- **Adversarial Attacks** are small imperceptible perturbation, malicously crafted to fool a deep learning-based classifier.



Background

Adversarial Robustness and Perceptually Aligned Gradients (PAG)

- **Adversarial Attacks** are small imperceptible perturbation, malicously crafted to fool a deep learning-based classifier.
- **Adversarial Robustness** requires models to be insensitive to small amounts of noise added to the input.

Background

Adversarial Robustness and Perceptually Aligned Gradients (PAG)

- **Adversarial Attacks**[1] are small imperceptible perturbation, malicously crafted to fool a deep learning-based classifier.
- **Adversarial Robustness** requires models to be insensitive to small amounts of noise added to the input.
  - A common technique for obtaining such classifiers is Adversarial Training[1,2]

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

1) Explaining and harnessing adversarial examples, Goodfellow et al., ICLR 2015
2) Towards Deep Learning Models Resistant to Adversarial Attacks, Madry et al., ICLR 2018

Background

Adversarial Robustness and Perceptually Aligned Gradients (PAG)

- The input-gradients of robust classifiers are semantically meaningful and more aligned with human perception

Background

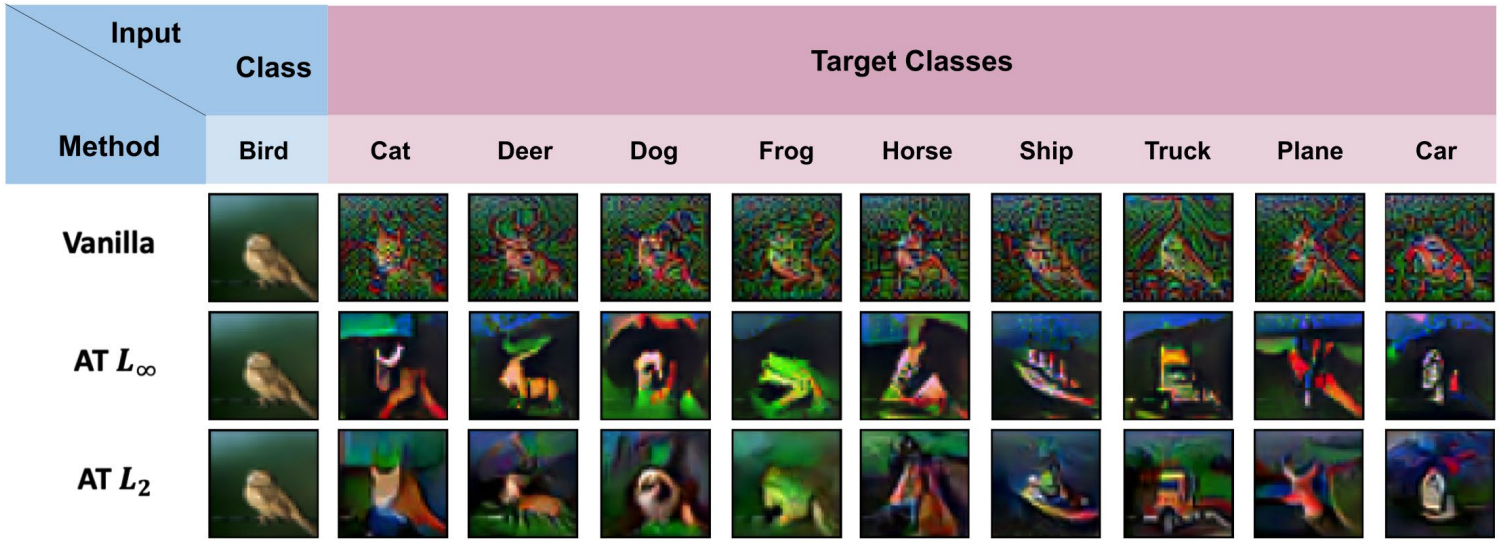
Adversarial Robustness and Perceptually Aligned Gradients (PAG)

- The input-gradients of robust classifiers are semantically meaningful and more aligned with human perception
  → As a result, strong targeted adversarial attacks on models with PAG leads to class related modifications

Background

Adversarial Robustness and Perceptually Aligned Gradients (PAG)

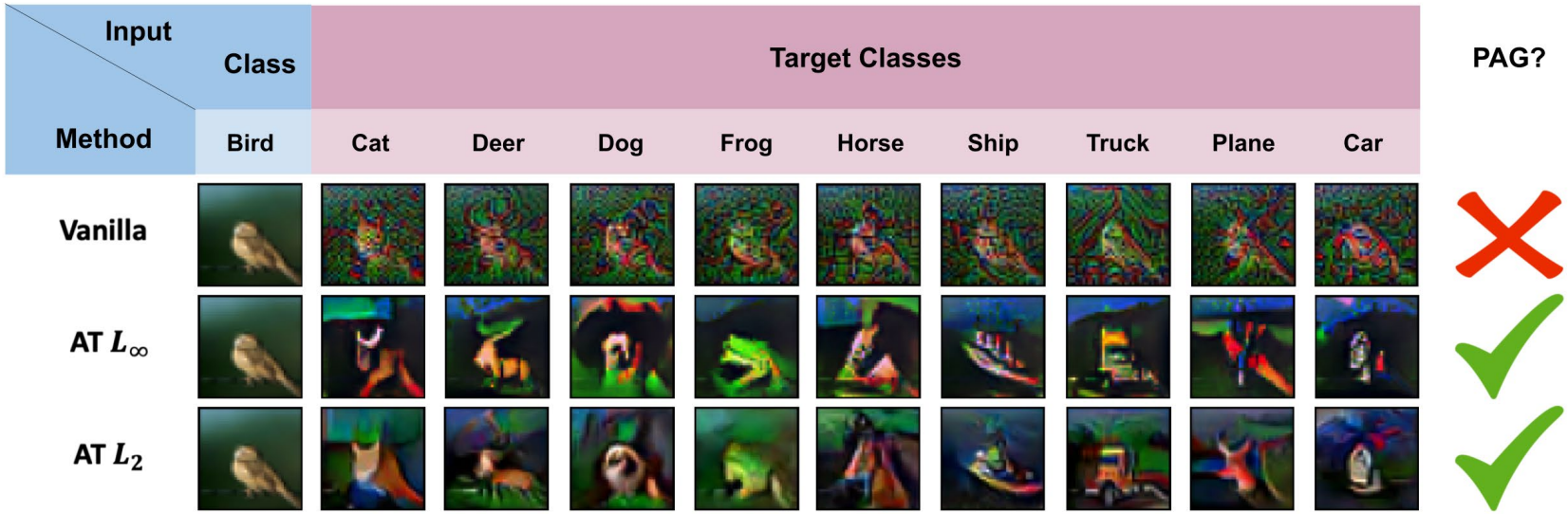
- The input-gradients of robust classifiers are semantically meaningful and more aligned with human perception
  → As a result, strong targeted adversarial attacks on models with PAG leads to class related modifications



Background

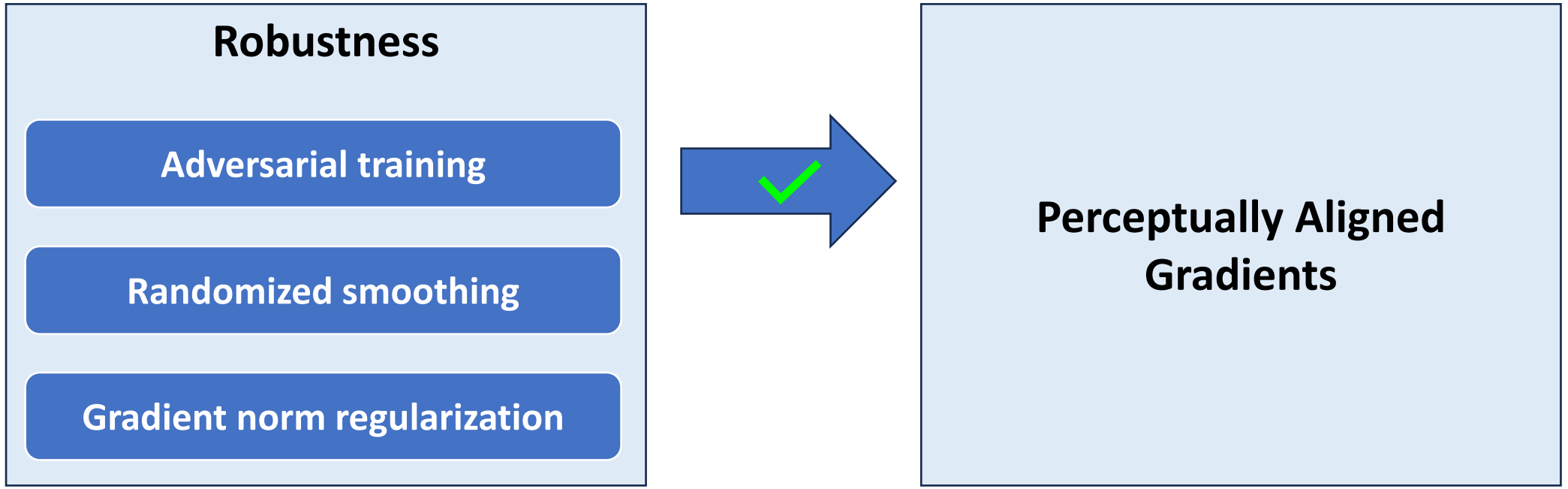
Adversarial Robustness and Perceptually Aligned Gradients (PAG)

- The input-gradients of robust classifiers are semantically meaningful and more aligned with human perception
  → As a result, strong targeted adversarial attacks on models with PAG leads to class related modifications



Motivation

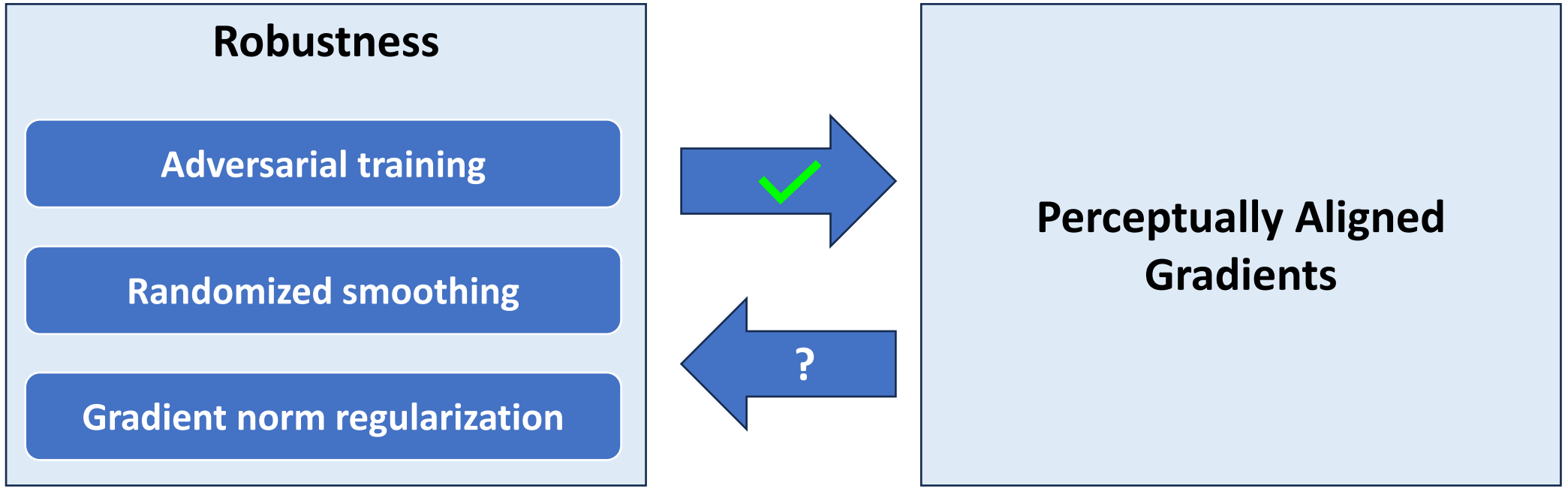
Perceptually Aligned Gradients (PAG) and Robustness



[1] Robustness may be at odds with accuracy. Tsipras *et al.*. ICLR 2019.
[2] Are perceptually-aligned gradients a general property of robust classifiers? Kaur *et al.* Arxiv.
[3] Rethinking the role of gradient-based attribution methods for model interpretability. Srinvas *et al.* ICLR 2021.

Motivation

Perceptually Aligned Gradients (PAG) and Robustness



[1] Robustness may be at odds with accuracy. Tsipras *et al.*. ICLR 2019.
[2] Are perceptually-aligned gradients a general property of robust classifiers? Kaur *et al.* Arxiv.
[3] Rethinking the role of gradient-based attribution methods for model interpretability. Srinvas *et al.* ICLR 2021.

Methodology

Perceptually Aligned Gradients (PAG) Training Method

Methodology

Perceptually Aligned Gradients (PAG) Training Method

- We develop an objective that induces PAG while disentangling it from adversarial training:

Methodology

Perceptually Aligned Gradients (PAG) Training Method

- We develop an objective that induces PAG while disentangling it from adversarial training:

$$\mathcal{L}_{total}(\boldsymbol{x}, y) = \overbrace{\mathcal{L}_{CE}(f_\theta(\boldsymbol{x}), y)}^{\textbf{Cross-entropy loss}} + \lambda \sum_{y_t \in \mathcal{C}} \mathcal{L}_{COS}\left(\nabla_x f_\theta(\boldsymbol{x})_{y_t}, g(\boldsymbol{x}, y_t)\right)$$

Methodology

Perceptually Aligned Gradients (PAG) Training Method

- We develop an objective that induces PAG while disentangling it from adversarial training:

$$\mathcal{L}_{total}(\boldsymbol{x}, y) = \overbrace{\mathcal{L}_{CE}(f_\theta(\boldsymbol{x}), y)}^{\textbf{Cross-entropy loss}} + \lambda \sum_{y_t \in \mathcal{C}} \underbrace{\mathcal{L}_{COS}\left(\nabla_x f_\theta(\boldsymbol{x})_{y_t}, g(\boldsymbol{x}, y_t)\right)}_{\textbf{PAG inducing term}}$$

Methodology

Perceptually Aligned Gradients (PAG) Training Method

- We develop an objective that induces PAG while disentangling it from adversarial training:

$$\mathcal{L}_{total}(\boldsymbol{x}, y) = \overbrace{\mathcal{L}_{CE}(f_\theta(\boldsymbol{x}), y)}^{\text{Cross-entropy loss}} + \underbrace{\lambda}_{\text{Alignment coeff.}} \sum_{y_t \in \mathcal{C}} \underbrace{\mathcal{L}_{COS}\left(\nabla_x f_\theta(\boldsymbol{x})_{y_t}, g(\boldsymbol{x}, y_t)\right)}_{\text{PAG inducing term}}$$

Methodology

Perceptually Aligned Gradients (PAG) Training Method

- We develop an objective that induces PAG while disentangling it from adversarial training:

$$\mathcal{L}_{total}(\boldsymbol{x}, y) = \overbrace{\mathcal{L}_{CE}(f_\theta(\boldsymbol{x}), y)}^{\text{Cross-entropy loss}} + \underbrace{\lambda}_{\text{Alignment coeff.}} \sum_{y_t \in \mathcal{C}} \underbrace{\mathcal{L}_{COS}\left(\nabla_x f_\theta(\boldsymbol{x})_{y_t}, g(\boldsymbol{x}, y_t)\right)}_{\text{PAG inducing term}}$$

- Avoiding circular reasoning by:

Methodology

Perceptually Aligned Gradients (PAG) Training Method

- We develop an objective that induces PAG while disentangling it from adversarial training:

$$\mathcal{L}_{total}(\boldsymbol{x}, y) = \overbrace{\mathcal{L}_{CE}(f_\theta(\boldsymbol{x}), y)}^{\text{Cross-entropy loss}} + \underbrace{\lambda}_{\text{Alignment coeff.}} \sum_{y_t \in \mathcal{C}} \underbrace{\mathcal{L}_{COS}\left(\nabla_x f_\theta(\boldsymbol{x})_{y_t}, g(\boldsymbol{x}, y_t)\right)}_{\text{PAG inducing term}}$$

- Avoiding circular reasoning by:
  - ✓ Not training on (adversarially) perturbed images

Methodology

Perceptually Aligned Gradients (PAG) Training Method

- We develop an objective that induces PAG while disentangling it from adversarial training:

$$\mathcal{L}_{total}(\boldsymbol{x}, y) = \overbrace{\mathcal{L}_{CE}(f_\theta(\boldsymbol{x}), y)}^{\text{Cross-entropy loss}} + \underbrace{\lambda}_{\text{Alignment coeff.}} \sum_{y_t \in \mathcal{C}} \underbrace{\mathcal{L}_{COS}\left(\nabla_x f_\theta(\boldsymbol{x})_{y_t}, g(\boldsymbol{x}, y_t)\right)}_{\text{PAG inducing term}}$$

- Avoiding circular reasoning by:
  - ✓ Not training on (adversarially) perturbed images
  - ✓ Not regularizing the input-gradients norm

Methodology

Perceptually Aligned Gradients (PAG) Training Method

- We develop an objective that induces PAG while disentangling it from adversarial training:

$$\mathcal{L}_{total}(\boldsymbol{x}, y) = \overbrace{\mathcal{L}_{CE}(f_\theta(\boldsymbol{x}), y)}^{\text{Cross-entropy loss}} + \underbrace{\lambda}_{\text{Alignment coeff.}} \sum_{y_t \in \mathcal{C}} \underbrace{\mathcal{L}_{COS}\left(\nabla_x f_\theta(\boldsymbol{x})_{y_t}, g(\boldsymbol{x}, y_t)\right)}_{\text{PAG inducing term}}$$

- Avoiding circular reasoning by:
  - ✓ Not training on (adversarially) perturbed images
  - ✓ Not regularizing the input-gradients norm

Requires access to $g(\boldsymbol{x}, y_t)$ - ground-truth Perceptually Aligned Gradients.

Methodology

Approximating Perceptually Aligned Gradients

Obtaining $g(x, y_t)$

Methodology

Approximating Perceptually Aligned Gradients

Obtaining $g(x, y_t)$

- The input-gradient of classification networks is $\nabla_x \log p(y|x)$

Methodology

Approximating Perceptually Aligned Gradients

Obtaining $g(x, y_t)$

- The input-gradient of classification networks is $\nabla_x \log p(y|x)$
- The conditional score-function, modeled by conditional diffusion models is $\nabla_{x_t} \log p(x_t|y)$

Methodology

Approximating Perceptually Aligned Gradients

Obtaining $g(x, y_t)$

- The input-gradient of classification networks is $\nabla_x \log p(y|x)$
- The conditional score-function, modeled by conditional diffusion models is $\nabla_{x_t} \log p(x_t|y)$
- Applying the Bayes rule $\rightarrow \nabla_{x_t} \log p(y|x_t) = \underbrace{\nabla_{x_t} \log p(x_t|y)}_{\text{Conditional diffusion}} - \overbrace{\nabla_{x_t} \log p(x_t)}^{\text{Unconditional diffusion}}$

Methodology

Approximating Perceptually Aligned Gradients

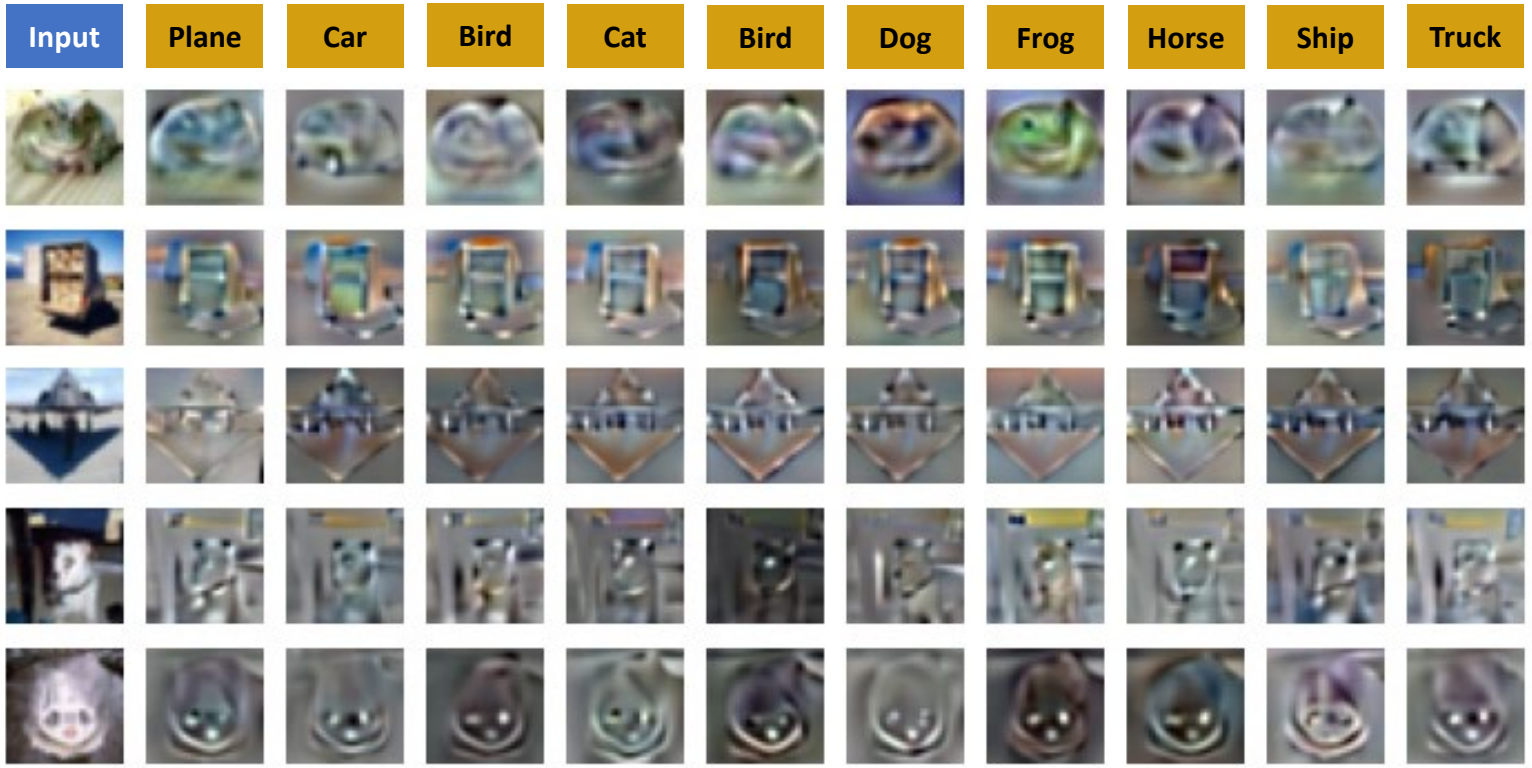
Obtaining $g(x, y_t)$

- The input-gradient of classification networks is $\nabla_x \log p(y|x)$
- The conditional score-function, modeled by conditional diffusion models is $\nabla_{x_t} \log p(x_t|y)$
- Applying the Bayes rule $\rightarrow \nabla_{x_t} \log p(y|x_t) = \underbrace{\nabla_{x_t} \log p(x_t|y)}_{\text{Conditional diffusion}} - \overbrace{\nabla_{x_t} \log p(x_t)}^{\text{Unconditional diffusion}}$

$\rightarrow$ Assuming $\log p(y|x) \approx \log p(y|x_t)$ for specific noise level t, approximate "ground-truth" PAG using **Score Based Gradients (SBG)**

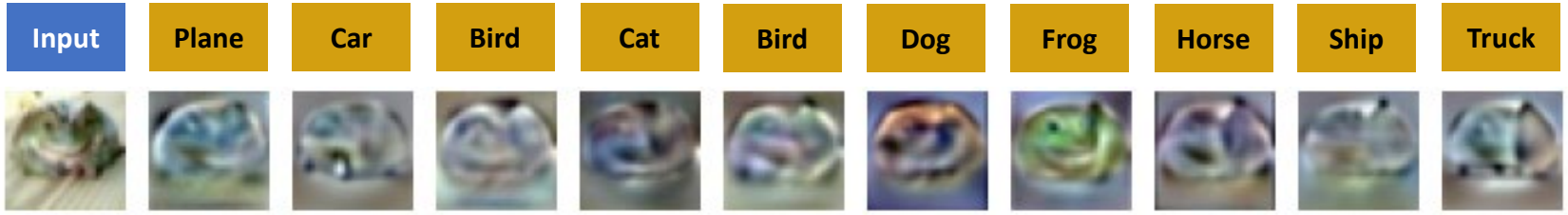
Methodology

Approximating Perceptually Aligned Gradients



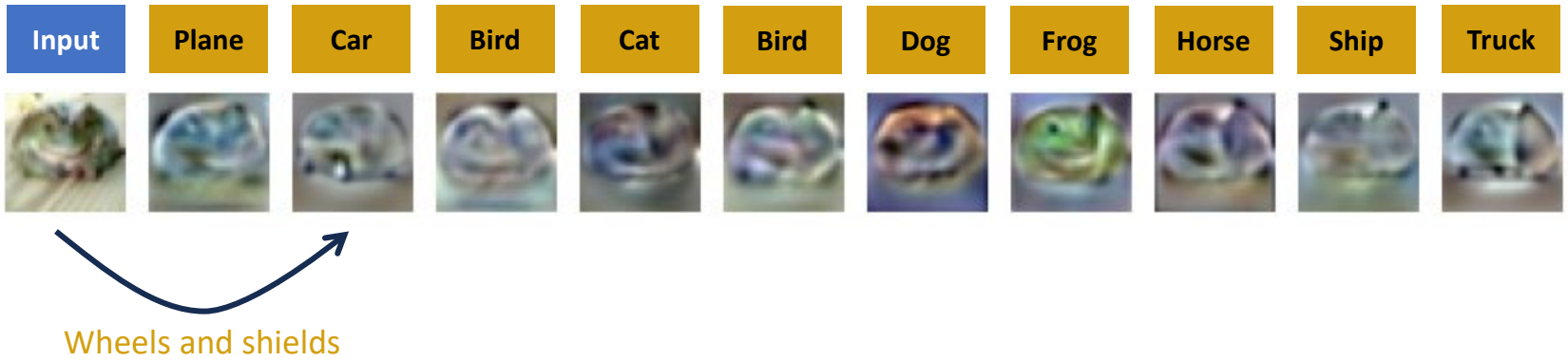
Methodology

Approximating Perceptually Aligned Gradients



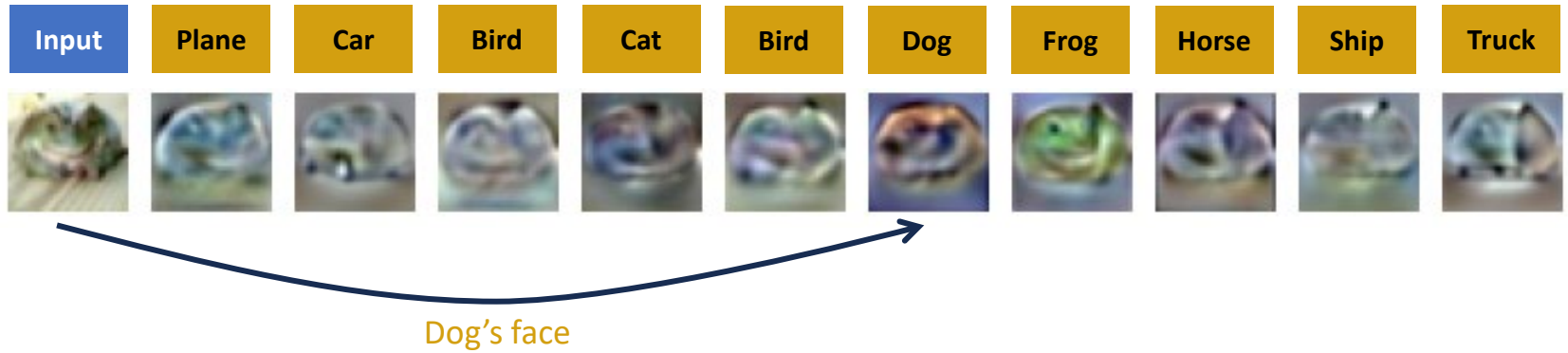
Methodology

Approximating Perceptually Aligned Gradients



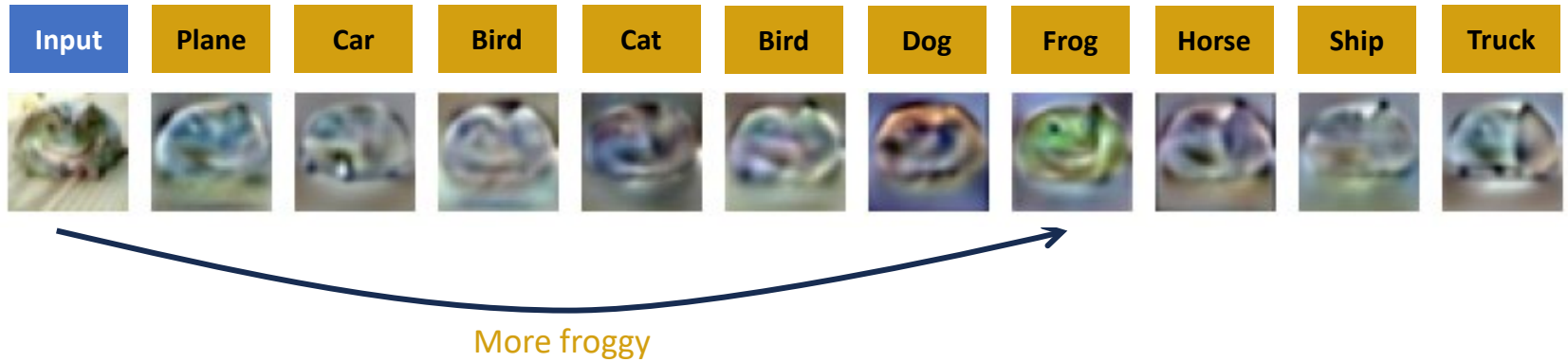
Methodology

Approximating Perceptually Aligned Gradients



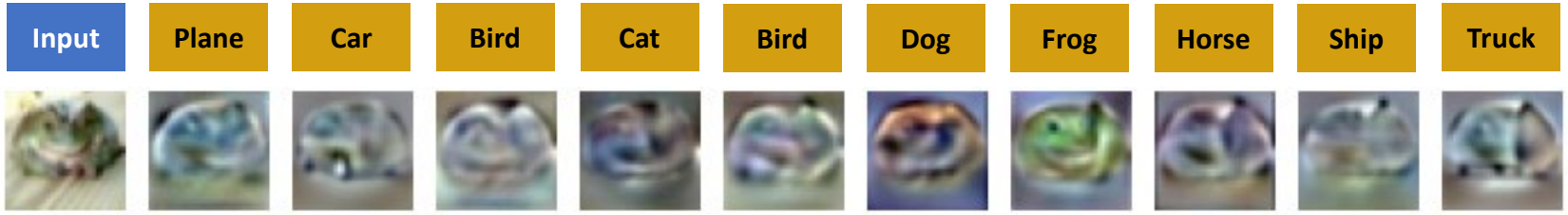
Methodology

Approximating Perceptually Aligned Gradients



Methodology

Approximating Perceptually Aligned Gradients



→ SBG performs meaningful perceptual modifications while maintaining the original image structure

Experiments

Do Perceptually Aligned Gradients Imply Robustness?

Approach

Experiments

Do Perceptually Aligned Gradients Imply Robustness?

Approach

1. Train classifiers with our proposed objective

Experiments

Do Perceptually Aligned Gradients Imply Robustness?

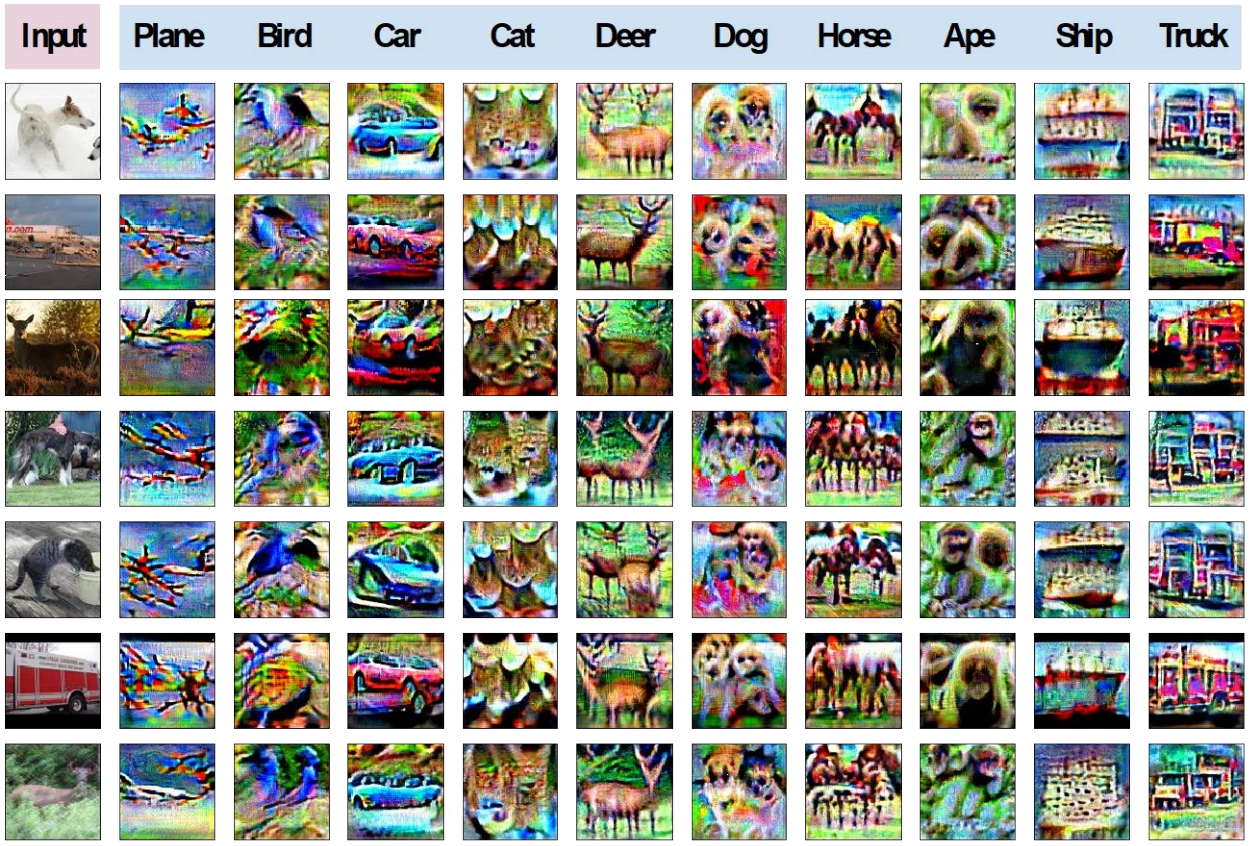
Approach

1. Train classifiers with our proposed objective

2. Qualitatively verify that such models possess PAG

CIFAR-10

STL



PAG visualizations using ResNet-18 on CIFAR-10 and STL.

Experiments

Do Perceptually Aligned Gradients Imply Robustness?

Approach

1. Train classifiers with our proposed objective
2. Qualitatively verify that such models possess PAG
3. Evaluate the adversarial robustness of such models

Experiments

Do Perceptually Aligned Gradients Imply Robustness?

Arch.	Method	Clean	AA L_2	AA L_∞
RN-18	Vanilla	**93.61**	00.00	00.00
	SBG	78.56	55.39	<u>23.97</u>

CIFAR-10 results using ResNet-18.

Experiments

Do Perceptually Aligned Gradients Imply Robustness?

Arch.	Method	Clean	AA L_2	AA L_∞
RN-18	Vanilla	**93.61**	00.00	00.00
	SBG	78.56	55.39	<u>23.97</u>
	AT L_∞	82.49	<u>56.57</u>	**37.59**
	AT L_2	<u>86.79</u>	**60.82**	19.63

CIFAR-10 results using ResNet-18.

Experiments

Do Perceptually Aligned Gradients Imply Robustness?

? Does this extends to different architectures?

Experiments

Do Perceptually Aligned Gradients Imply Robustness?

✔ Does this extends to different architectures?

Arch.	Method	Clean	AA L_2	AA L_∞
ViT	Vanilla	<u>80.51</u>	00.87	00.01
	SBG	**81.28**	**57.80**	<u>22.85</u>
	AT L_∞	62.20	42.80	**24.62**
	AT L_2	72.81	<u>42.99</u>	08.13

CIFAR-10 results using ViT.

Experiments

Do Perceptually Aligned Gradients Imply Robustness?

? Does this extends to low-data regimes?

Experiments

Do Perceptually Aligned Gradients Imply Robustness?

✔ Does this extends to low-data regimes?

Method	Clean	AA L_2	AA L_∞
Vanilla	**82.60**	00.00	00.00
SBG	<u>74.79</u>	**65.96**	**43.53**
AT L_∞	54.90	46.33	28.30
AT L_2	54.99	46.04	23.33

STL results using ResNet-18.

Experiments

Do Perceptually Aligned Gradients Imply Robustness?

? Does our method extends to datasets with more classes?

Experiments

Do Perceptually Aligned Gradients Imply Robustness?

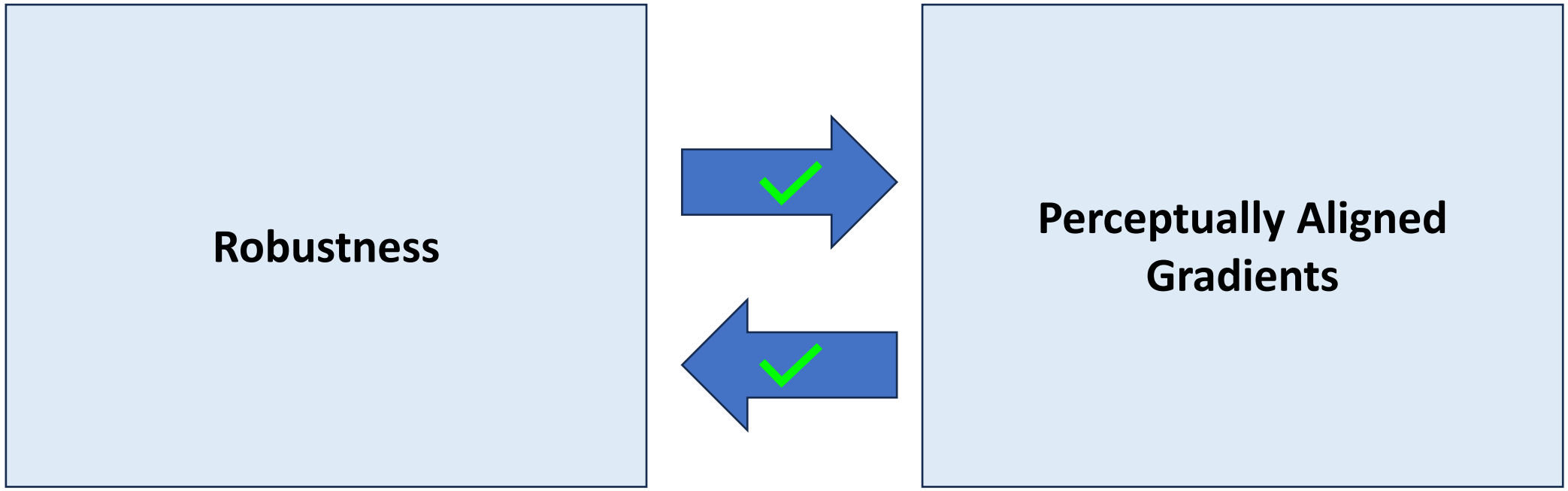
✔ Does our method extends to datasets with more classes?

Method	Clean	AA L_2	AA L_∞
Vanilla	**74.36**	00.00	00.00
SBG	55.94	<u>29.25</u>	<u>08.24</u>
AT L_∞	52.92	23.61	**14.63**
AT L_2	58.05	**30.51**	08.03

CIFAR-100 results using ResNet-18.

Conclusions

Do Perceptually Aligned Gradients Imply Robustness?



Arxiv

Code