Adversarial Attacks are Reversible with Natural Supervision

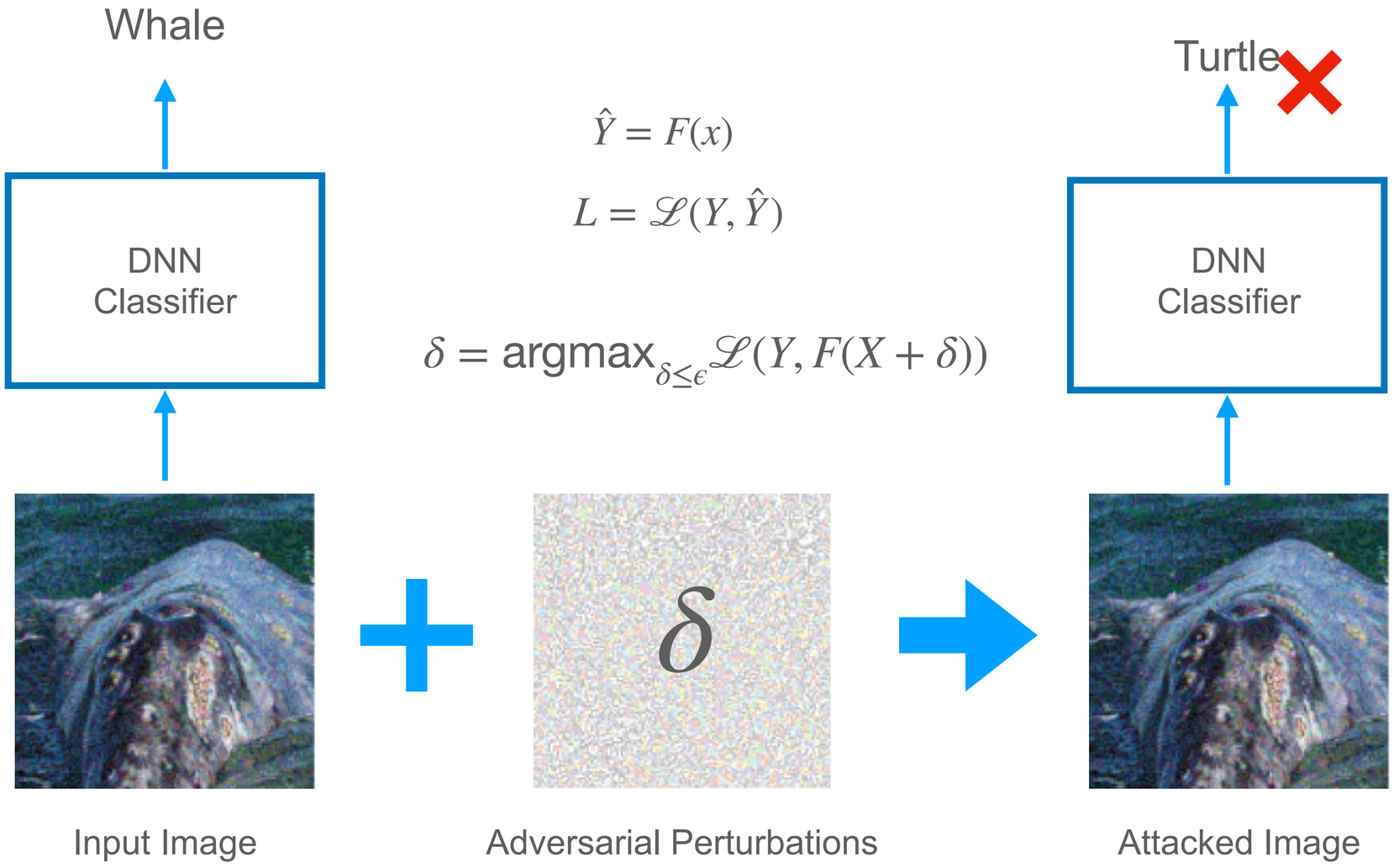
Chengzhi Mao, Mia Chiquier, Hao Wang*, Junfeng Yang, Carl Vondrick





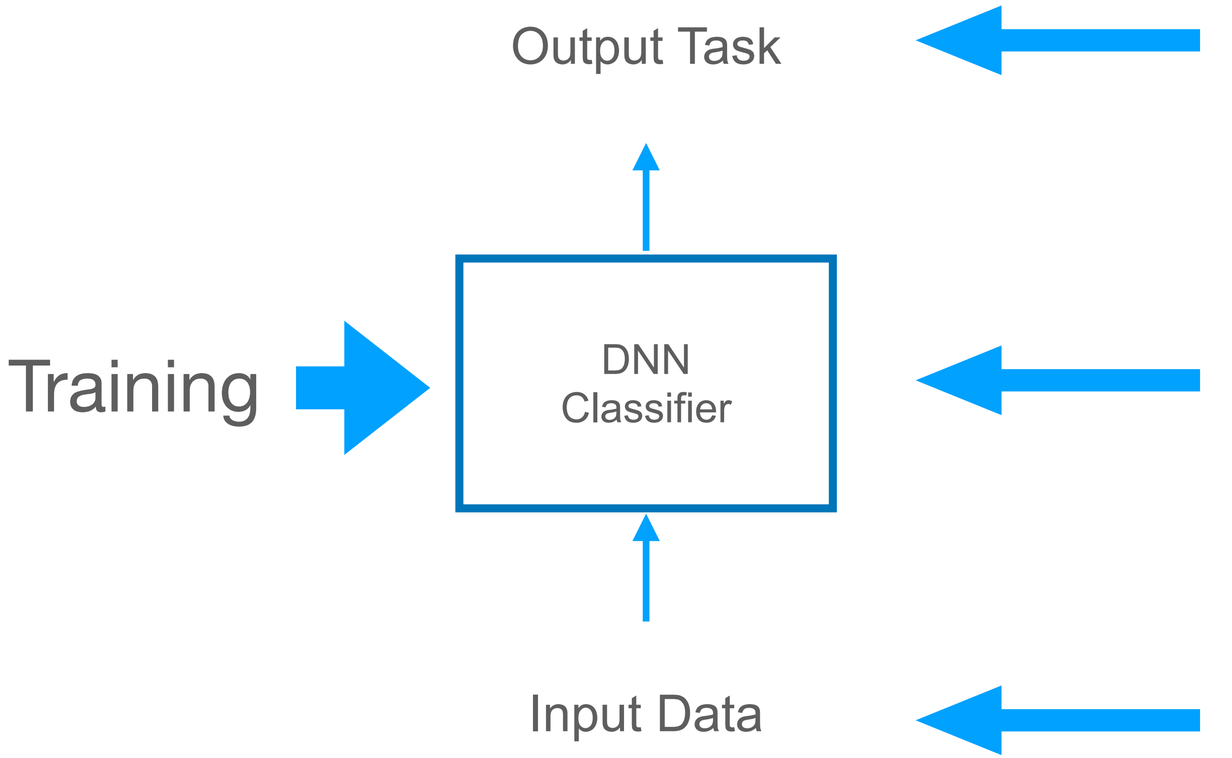
Columbia University, Rutgers University*

Deep Networks are Vulnerable under Adversarial Attacks



Moosavi-Dezfooli et al.

Existing defense focus on the training algorithm



Multiple Tasks [3]

Regularization [2]

$$\sum_i^N \mathcal{L}_{trip}(\mathbf{x}_a^{(i)}, \mathbf{x}_p^{(i)}, \mathbf{x}_n^{(i)}) = \sum_i^N [D(h(\mathbf{x}_a^{(i)}), h(\mathbf{x}_p^{(i)})) - D(h(\mathbf{x}_a^{(i)}), h(\mathbf{x}_n^{(i)})) + \alpha]_+$$

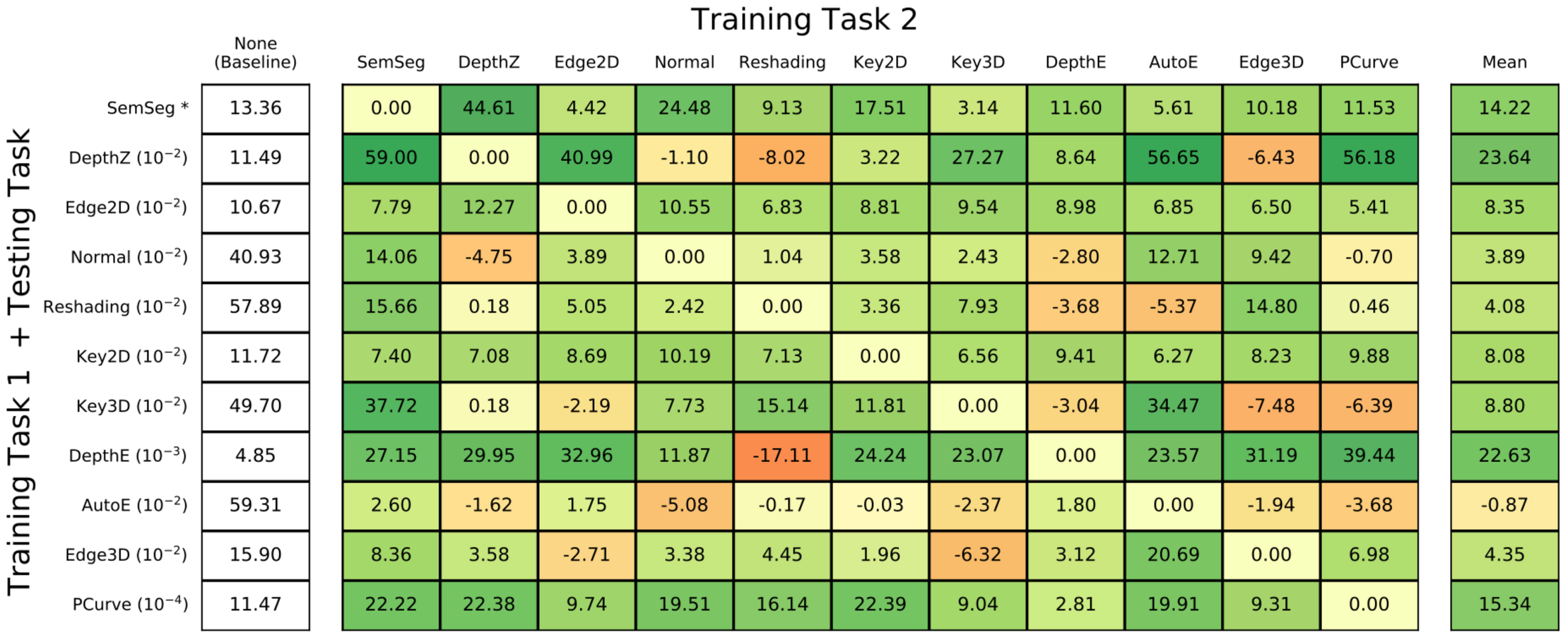
Adversarial Training [1]

$$\delta = \text{argmax}_{\delta \leq \epsilon} \mathscr{L}(Y, F_\theta(X + \delta))$$

$$\theta = \text{argmin}_\theta(\mathscr{L}(Y, F_\theta(X + \delta)))$$

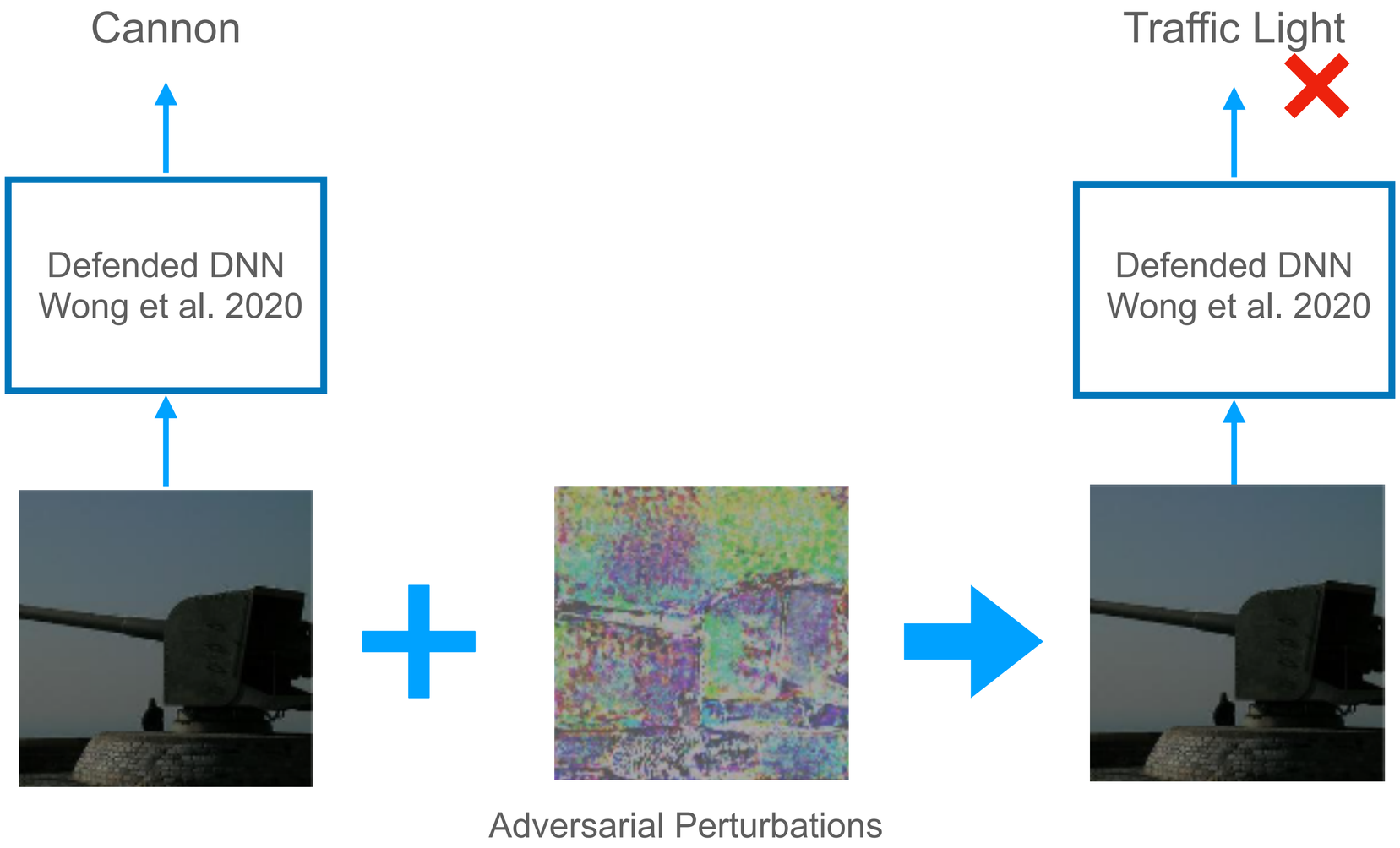
[1] Madry et al. Towards deep learning models resistant to adversarial attacks
[2] Mao et al. Metric Learning for Adversarial Robustness.
[3] Mao et al. Multitask learning strengthens adversarial robustness.

Additional tasks during training secure the model

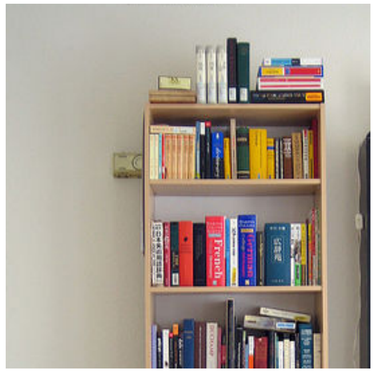


(a) Performance Under Attack

Training time defense is not enough



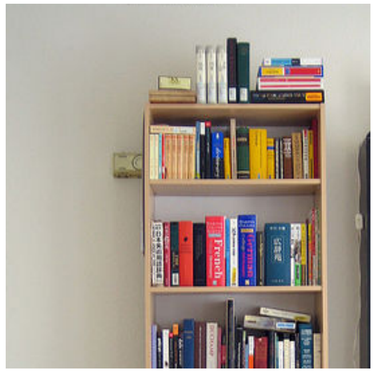
Natural Images contain intrinsic structure





Natural Images contain intrinsic structure

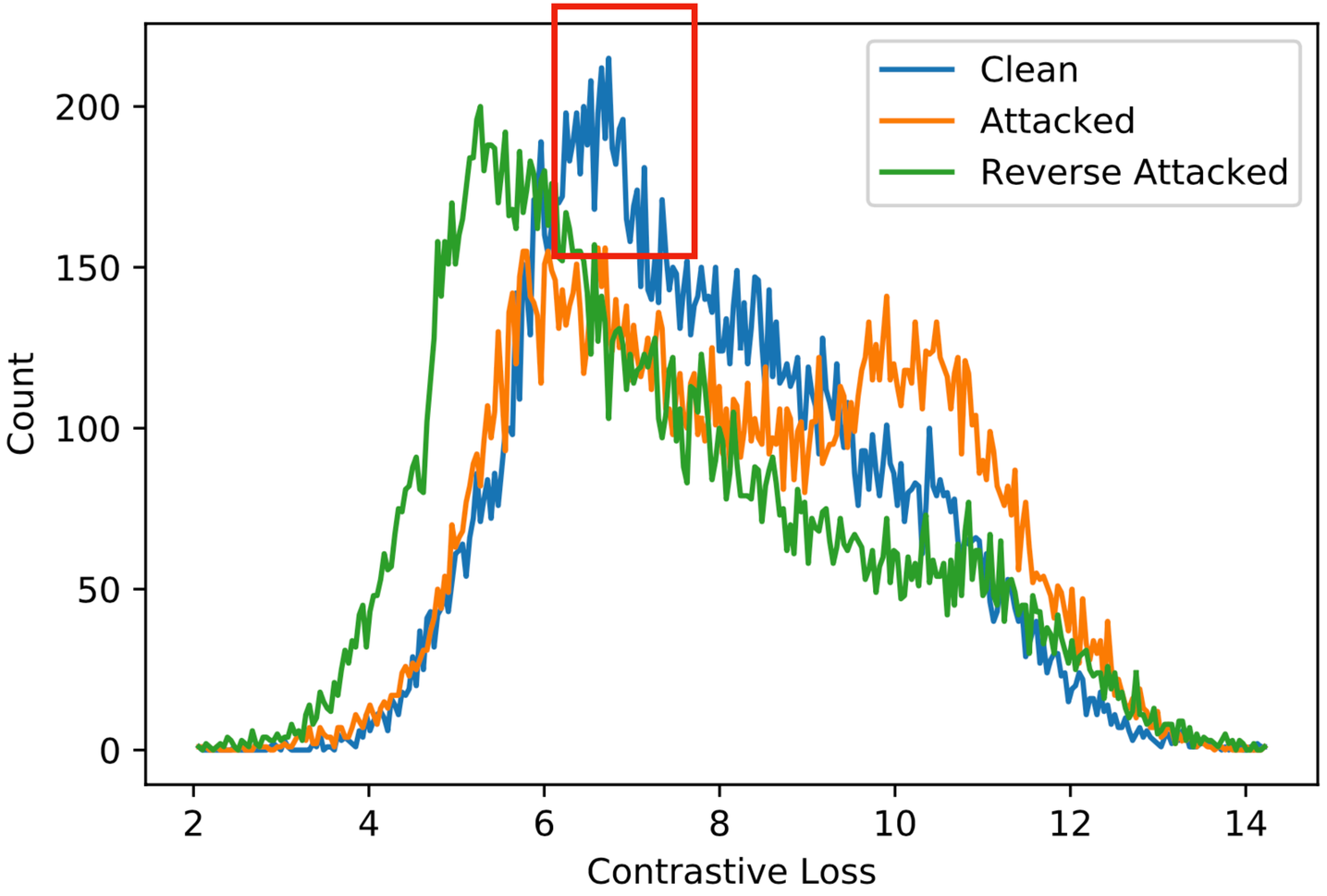




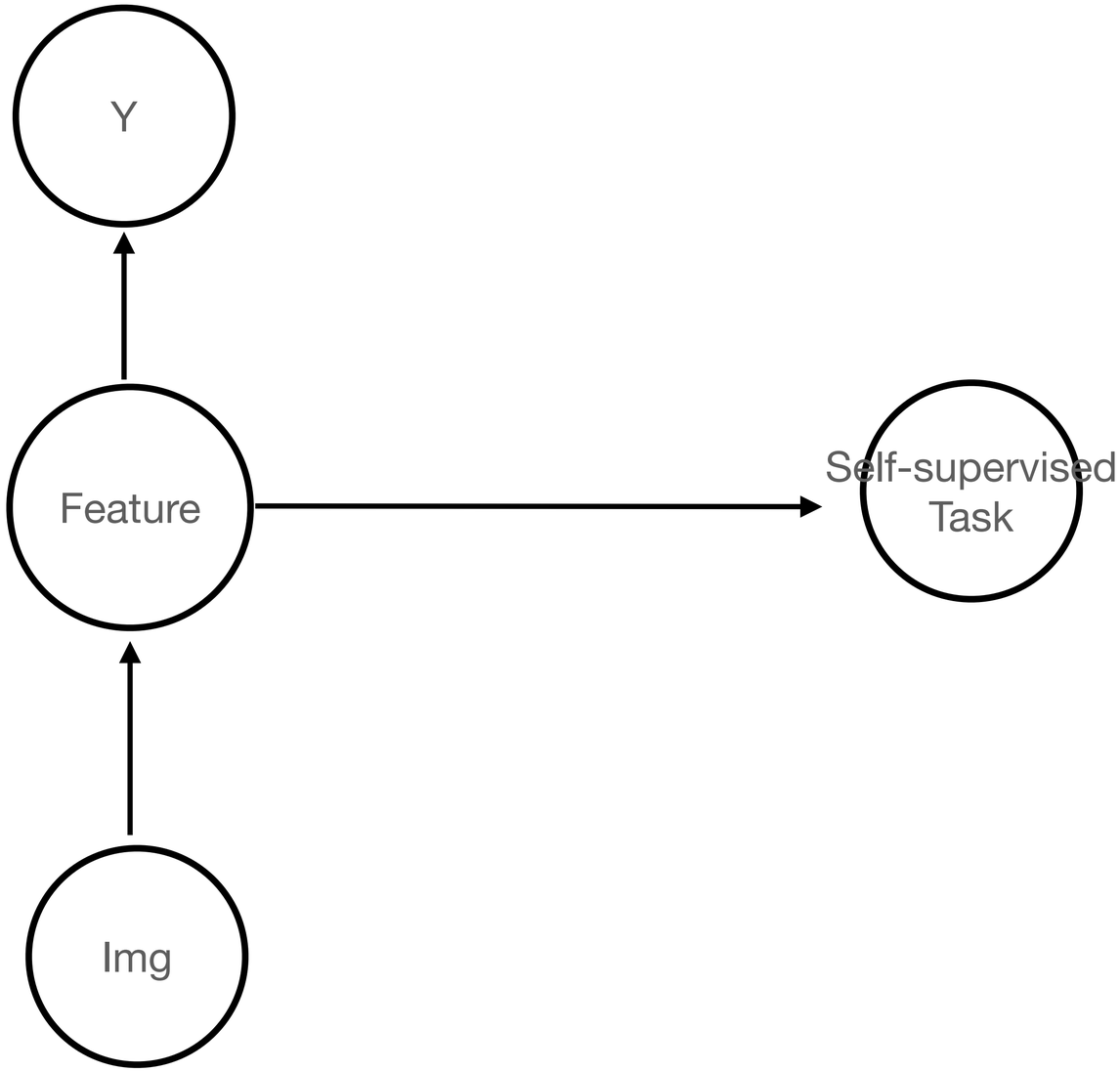
Natural Images contain intrinsic structure



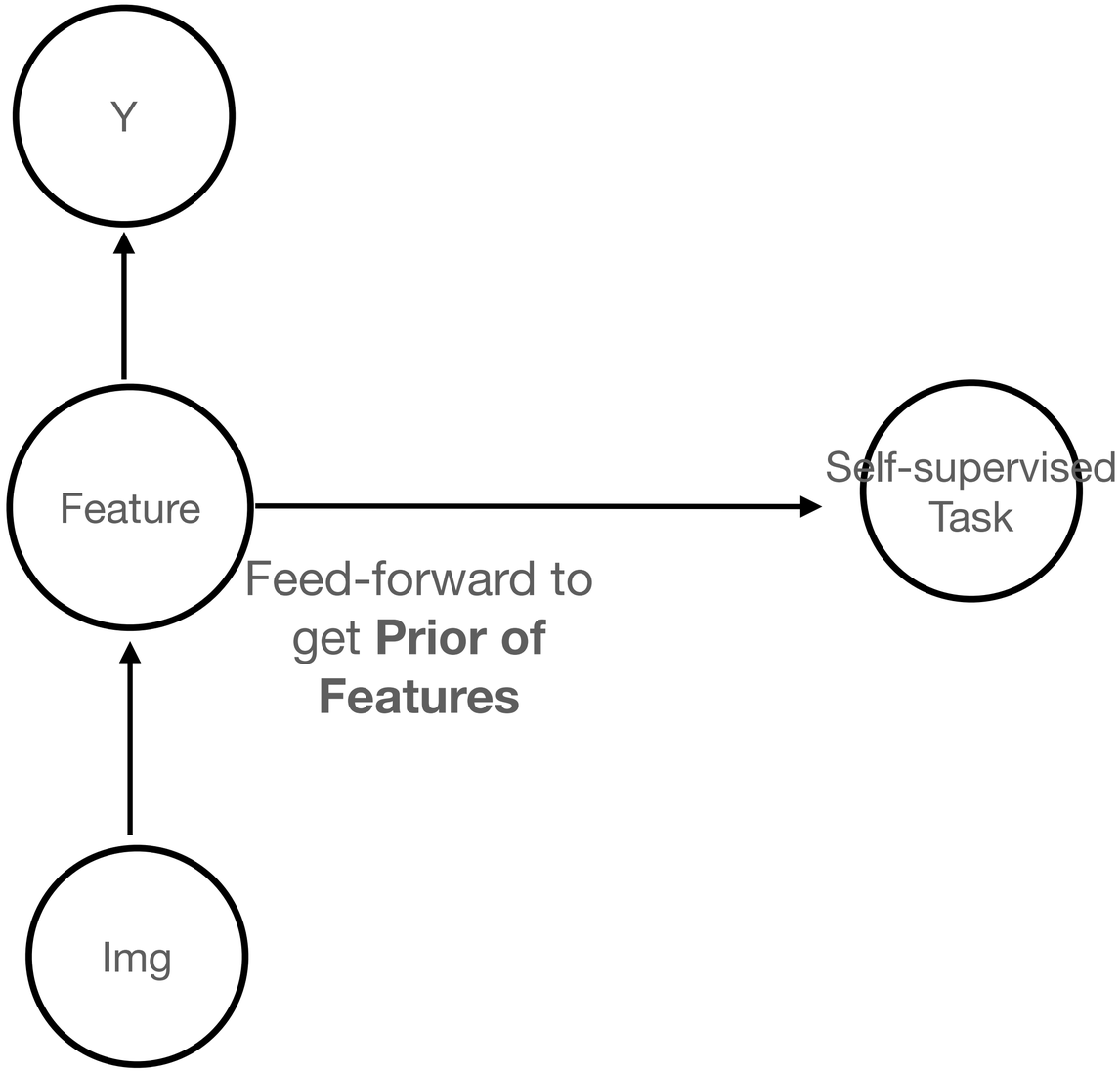

Adversarial Attack also damages the incidental structure of image



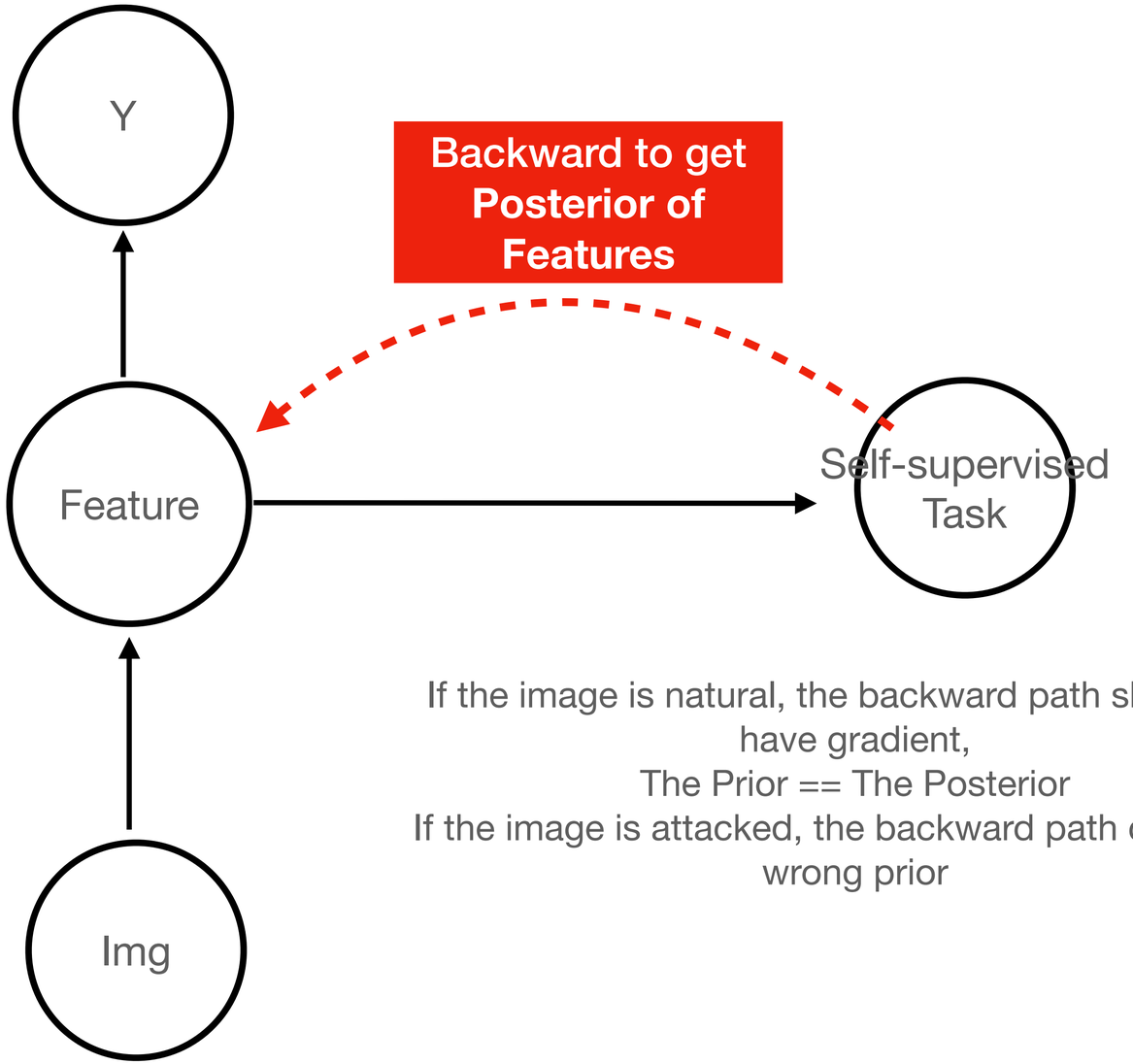
Graphical Model



Graphical Model

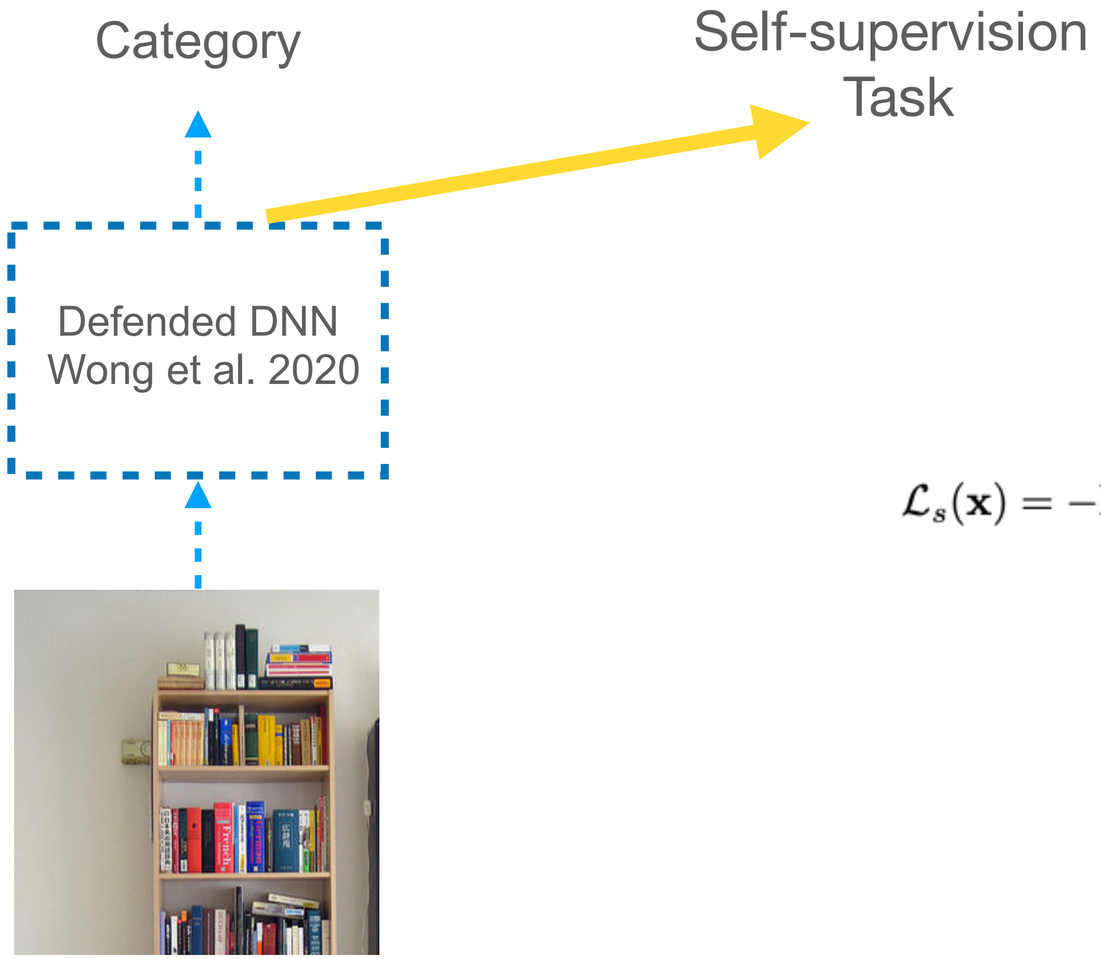


Graphical Model



If the image is natural, the backward path should not have gradient,
The Prior == The Posterior
If the image is attacked, the backward path can fix the wrong prior

Algorithm



Contrastive Loss

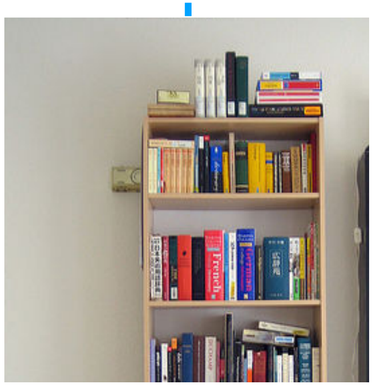
$$\mathcal{L}_s(\mathbf{x}) = -\mathbb{E}_{i,j}\left[\mathbf{y}_{ij}^{(s)} \log \frac{\exp(\cos(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_k \exp(\cos(\mathbf{z}_i, \mathbf{z}_k)/\tau)}\right]$$

Contrastive Loss for Self-supervision

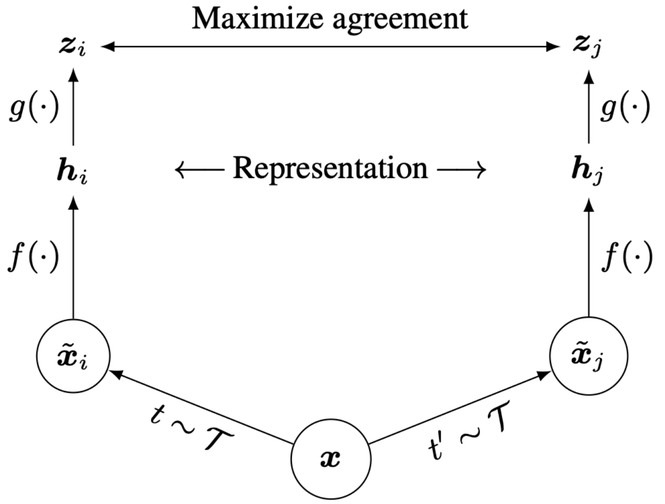
Category

Self-supervision Task

Defended DNN
Wong et al. 2020



$$\mathcal{L}_s(\mathbf{x}) = -\mathbb{E}_{i,j}\left[\mathbf{y}_{ij}^{(s)} \log \frac{\exp(\cos(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_k \exp(\cos(\mathbf{z}_i, \mathbf{z}_k)/\tau)}\right]$$

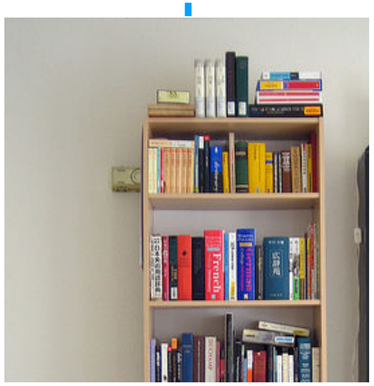


Contrastive Loss for Self-supervision

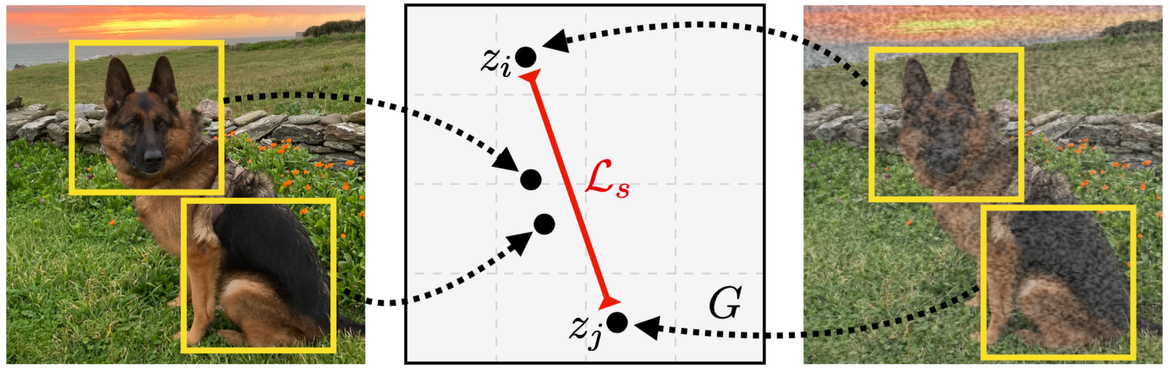
Category

Self-supervision Task

Defended DNN
Wong et al. 2020



$$\mathcal{L}_s(\mathbf{x}) = -\mathbb{E}_{i,j}\left[\mathbf{y}_{ij}^{(s)} \log \frac{\exp(\cos(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_k \exp(\cos(\mathbf{z}_i, \mathbf{z}_k)/\tau)}\right]$$

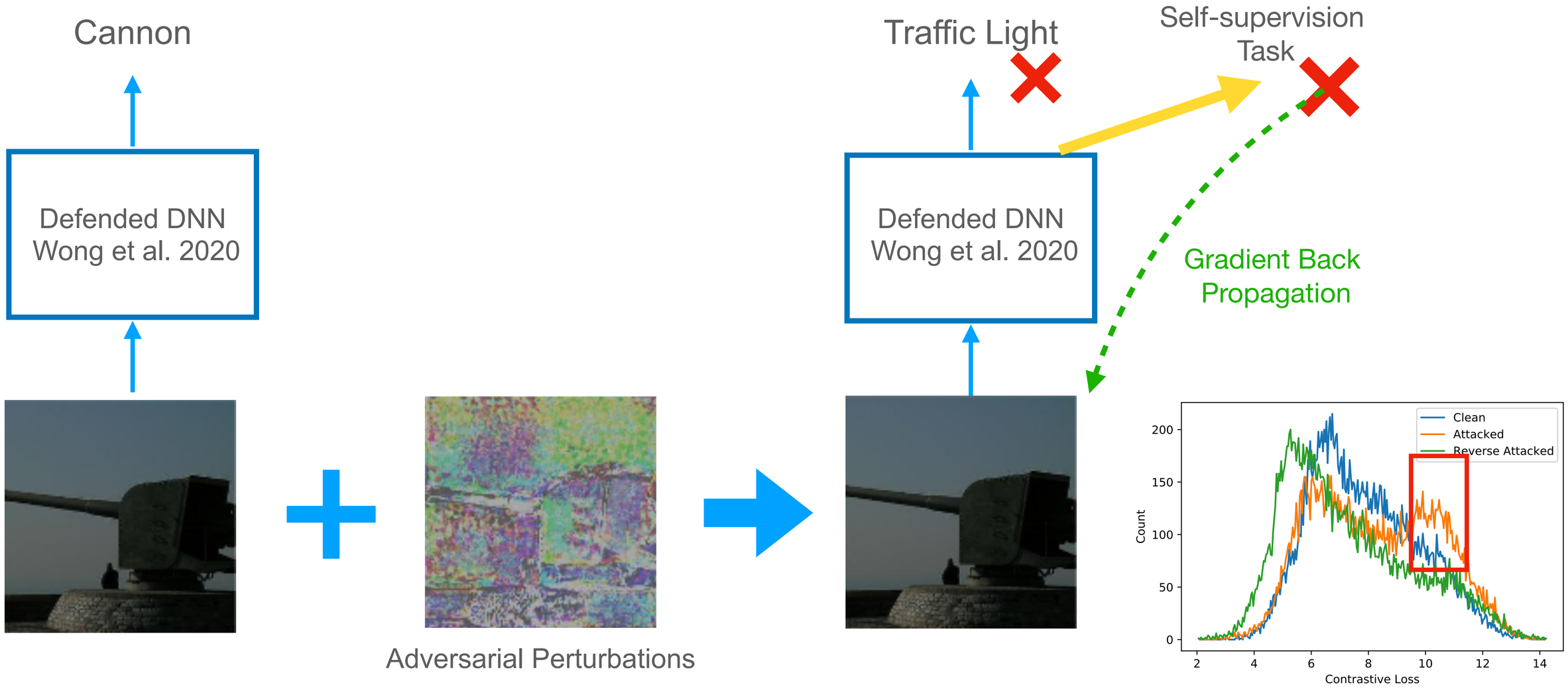


Original Image
x

Contrastive Embeddings

Attacked Image
x_a

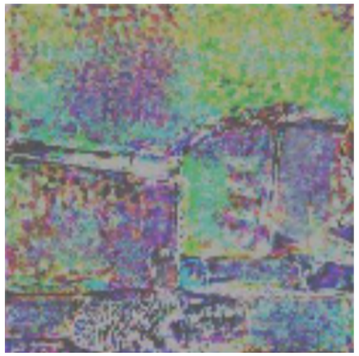
Our Method


Cannon
Defended DNN
Wong et al. 2020
Adversarial Perturbations
Traffic Light
Self-supervision
Task
Defended DNN
Wong et al. 2020
Gradient Back
Propagation
Clean
Attacked
Reverse Attacked
Count
Contrastive Loss
0
50
100
150
200
2
4
6
8
10
12
14

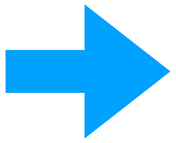
Our Method

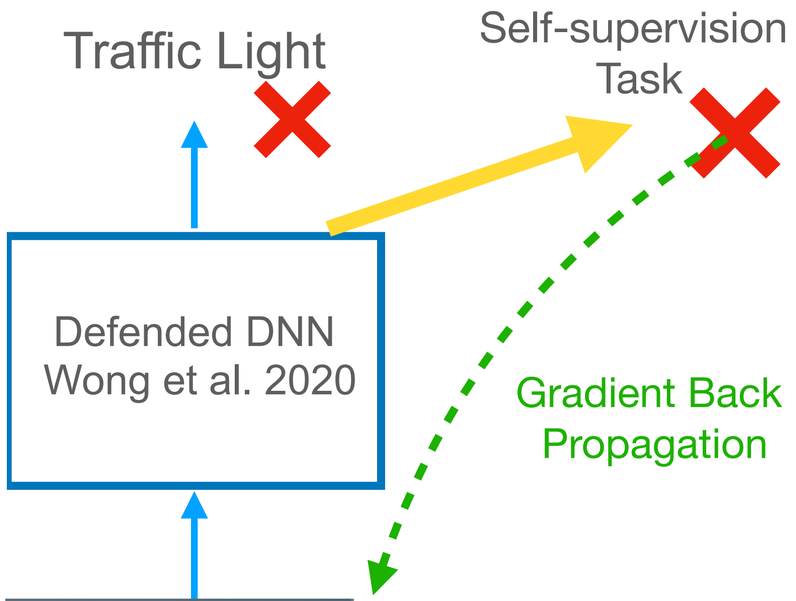
Algorithm 1 Self-supervised Reverse Attack

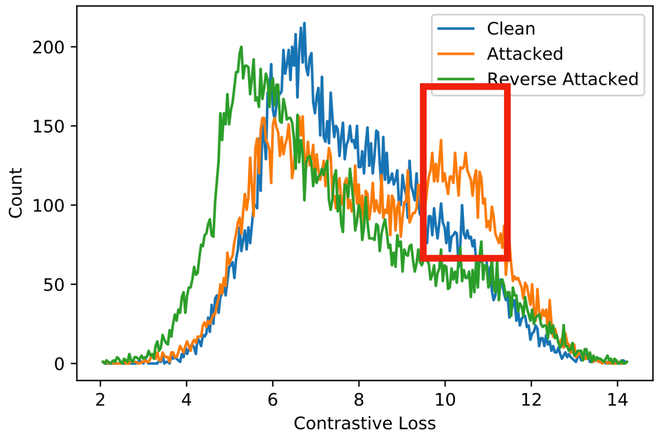
Input: Potentially attacked image $\mathbf{x}$, step size η, number of iterations K, a classifier F, reverse attack bound ϵ_v, and self-supervised loss function $\mathcal{L}_s$.
Output: Class prediction $\hat{y}$
Inference:
$\mathbf{x}' \leftarrow \mathbf{x} + \mathbf{n}$, where $\mathbf{n}$ is the initial random noise
for $k = 1, ..., K$ **do**
  $\mathbf{x}' \leftarrow \mathbf{x}' - \eta \nabla_{\mathbf{x}'} \mathcal{L}_s(\mathbf{x}')$
  $\mathbf{x}' \leftarrow \Pi_{(\mathbf{x}, \epsilon_v)} \mathbf{x}'$, which projects the image back into the bounded region.
end for
Predict the final output by $\hat{y} = F(\mathbf{x}')$



Adversarial Perturbations



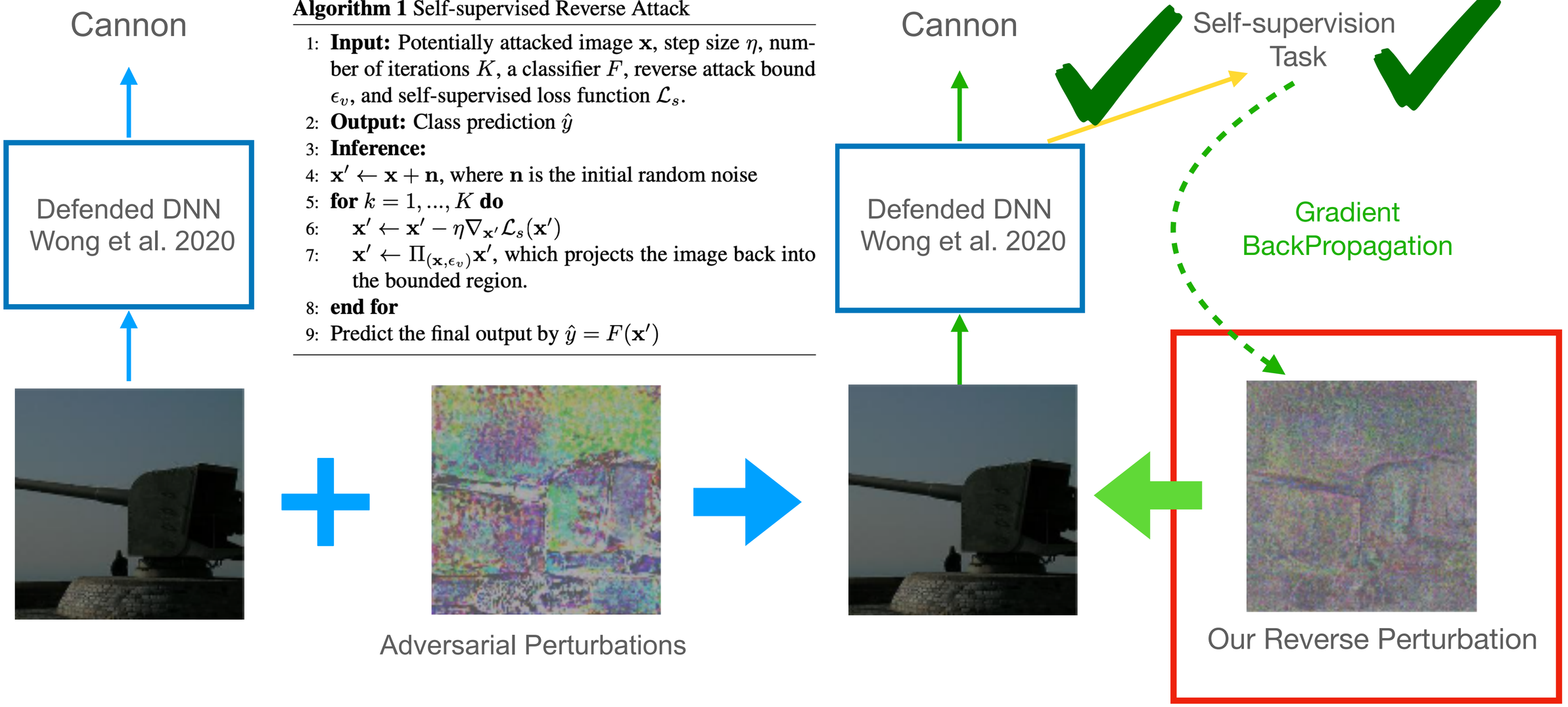




Our Method

Algorithm 1 Self-supervised Reverse Attack

1: **Input:** Potentially attacked image $\mathbf{x}$, step size η, number of iterations K, a classifier F, reverse attack bound ϵ_v, and self-supervised loss function $\mathcal{L}_s$.
2: **Output:** Class prediction $\hat{y}$
3: **Inference:**
4: $\mathbf{x}' \leftarrow \mathbf{x} + \mathbf{n}$, where $\mathbf{n}$ is the initial random noise
5: **for** $k = 1, ..., K$ **do**
6: $\quad \mathbf{x}' \leftarrow \mathbf{x}' - \eta \nabla_{\mathbf{x}'} \mathcal{L}_s(\mathbf{x}')$
7: $\quad \mathbf{x}' \leftarrow \Pi_{(\mathbf{x}, \epsilon_v)} \mathbf{x}'$, which projects the image back into the bounded region.
8: **end for**
9: Predict the final output by $\hat{y} = F(\mathbf{x}')$

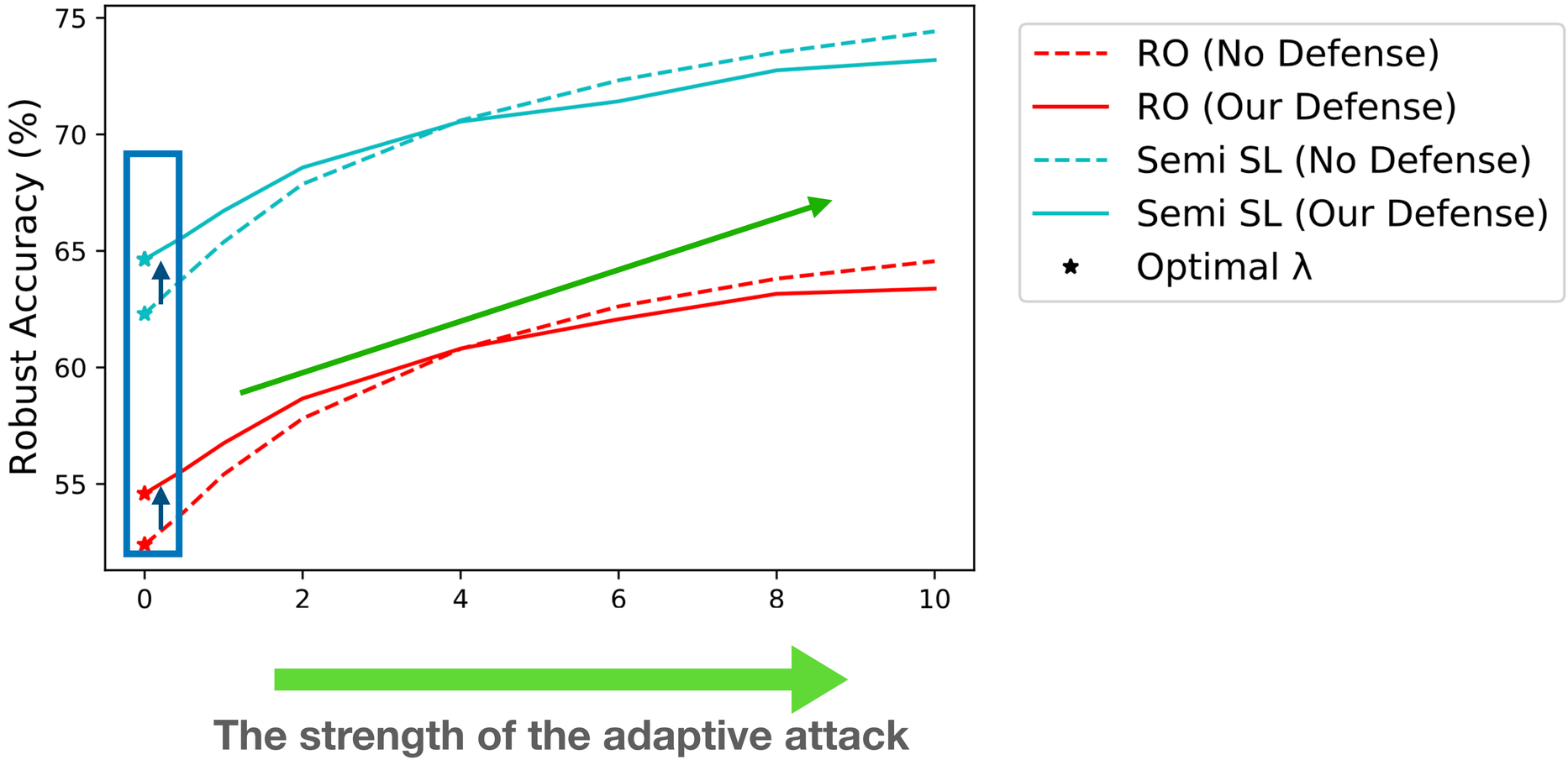


Adaptive attack: Lose-lose situation for attacker

If the attack ignores our defense, our defense improves accuracy

If the attacker adapts to our defense, it will hurt their attack and increase our robust accuracy.

This makes sense which means the attacker help us find a attack that also produces a good prior



Theoretical Guarantee

Our approach can improve the classification accuracy upper bound.

Lemma 1. *The standard classifier under adversarial attack is equivalent to predicting with* $P(\mathbf{Y}|\mathbf{X}=\mathbf{x}_a, \mathbf{Y}^{(s)}=\mathbf{y}_a^{(s)})$*, and our approach is equivalent to predicting with* $P(\mathbf{Y}|\mathbf{X}=\mathbf{x}_a, \mathbf{Y}^{(s)}=\mathbf{y}^{(s)})$.

Theorem 1. *Assume the base classifier operates better than chance and instances in the dataset are uniformly distributed over* n *categories. Let the prediction accuracy bounds be* $P(\mathbf{Y}|\mathbf{Y}_a^{(s)}, \mathbf{X}_a) \in [b_1, c_1]$ *and* $P(\mathbf{Y}|\mathbf{Y}^{(s)}, \mathbf{X}_a) \in [b_2, c_2]$*. If the conditional mutual information* $I(\mathbf{Y}; \mathbf{Y}^{(s)}|\mathbf{X}_a) > 0$*, we have* $b_2 \geq b_1$ *and* $c_2 > c_1$*, which means our approach strictly improves the bound for classification accuracy.*

Theoretical Guarantee

Lemma 1. *The standard classifier under adversarial attack is equivalent to predicting with* $P(\mathbf{Y}|\mathbf{X}=\mathbf{x}_a, \mathbf{Y}^{(s)}=\mathbf{y}_a^{(s)})$*, and our approach is equivalent to predicting with* $P(\mathbf{Y}|\mathbf{X}=\mathbf{x}_a, \mathbf{Y}^{(s)}=\mathbf{y}^{(s)})$.

Proof. For the standard classifier under attack, we know that $P(\mathbf{Y}^{(s)}=\mathbf{y}_a^{(s)}|\mathbf{X}=\mathbf{x}_a)=1$. Thus we know the standard classifier under adversarial attack is equivalent to

$$\begin{aligned} P(\mathbf{Y}|\mathbf{X}=\mathbf{x}_a) &= \sum_{\mathbf{Y}^{(s)}} P(\mathbf{Y}^{(s)}|\mathbf{X}=\mathbf{x}_a) P(\mathbf{Y}|\mathbf{Y}^{(s)}, \mathbf{X}=x_a) \\ &= P(\mathbf{Y}|\mathbf{Y}^{(s)}=\mathbf{y}_a^{(s)}, \mathbf{X}=\mathbf{x}_a). \end{aligned}$$

Our algorithm finds a new input image $\mathbf{x}_{\max}^{(n)}$ that

$$\begin{aligned} &\operatorname*{argmax}_{\mathbf{x}^{(n)}} P(\mathbf{X}^{(n)}=\mathbf{x}^{(n)}|\mathbf{X}=\mathbf{x}_a) P(\mathbf{Y}^{(s)}=\mathbf{y}^{(s)}|\mathbf{X}^{(n)}=\mathbf{x}^{(n)}) \\ &= \operatorname*{argmax}_{\mathbf{x}^{(n)}} P(\mathbf{X}^{(x)}=\mathbf{x}^{(n)}|\mathbf{X}=\mathbf{x}_a, \mathbf{Y}^{(s)}=\mathbf{y}^{(s)}). \end{aligned}$$

Our algorithm first estimate $\mathbf{x}_{\max}^{(n)}$ with adversarial image $\mathbf{x}_a$ and self-supervised label $\mathbf{y}^{(s)}$. We then predict the label $\mathbf{Y}$ using our new image $\mathbf{x}_{\max}^{(n)}$. Thus, our approach in fact estimates $P(\mathbf{Y}|\mathbf{X}^{(n)}=\mathbf{x}_{\max}^{(n)}) P(\mathbf{X}^{(n)}=\mathbf{x}_{\max}^{(n)}|\mathbf{X}=\mathbf{x}_a, \mathbf{Y}^{(s)}=\mathbf{y}^{(s)})$. Note the following holds:

$$P(\mathbf{Y}|\mathbf{X}=\mathbf{x}_a, \mathbf{Y}^{(s)}=\mathbf{y}^{(s)}) \tag{9}$$

$$= \sum_{\mathbf{x}^{(n)}} P(\mathbf{Y}|\mathbf{x}^{(n)}) P(\mathbf{x}^{(n)}|\mathbf{X}=\mathbf{x}_a, \mathbf{Y}^{(s)}=\mathbf{y}^{(s)}) \tag{10}$$

$$\approx P(\mathbf{Y}|\mathbf{X}^{(n)}=\mathbf{x}_{\max}^{(n)}) P(\mathbf{X}^{(n)}=\mathbf{x}_{\max}^{(n)}|\mathbf{X}=\mathbf{x}_a, \mathbf{Y}^{(s)}=\mathbf{y}^{(s)}) \tag{11}$$

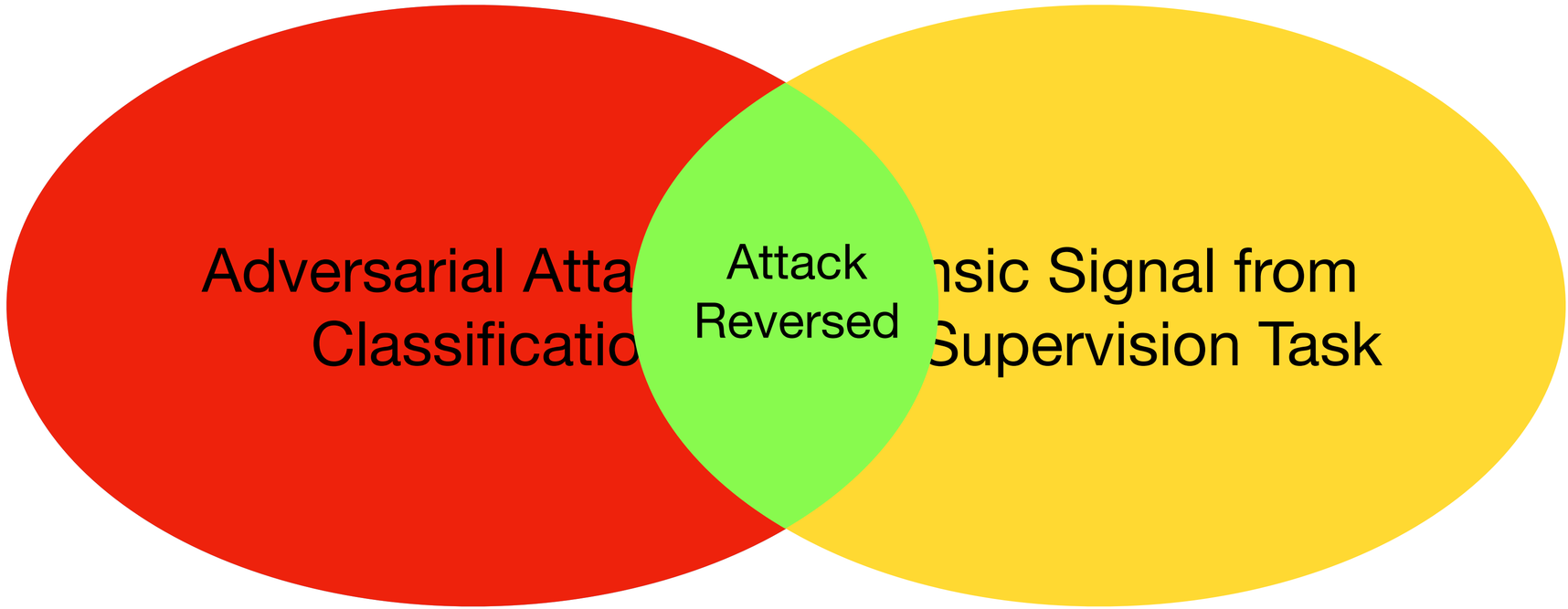
Thus our approach is equivalent to estimating $P(\mathbf{Y}|\mathbf{X}=\mathbf{x}_a, \mathbf{Y}^{(s)}=\mathbf{y}^{(s)})$. □

Theoretical Guarantee

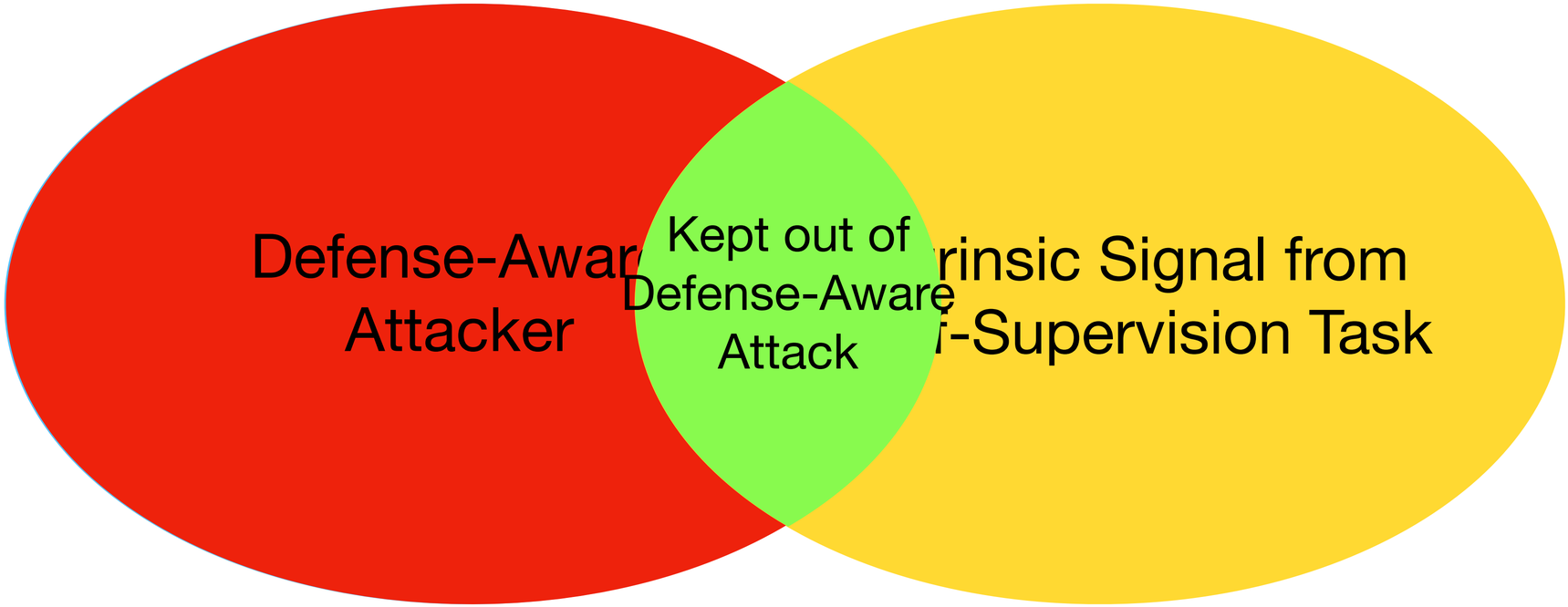
Theorem 1. *Assume the base classifier operates better than chance and instances in the dataset are uniformly distributed over n categories. Let the prediction accuracy bounds be $P(\mathbf{Y}|\mathbf{Y}_a^{(s)}, \mathbf{X}_a) \in [b_1, c_1]$ and $P(\mathbf{Y}|\mathbf{Y}^{(s)}, \mathbf{X}_a) \in [b_2, c_2]$. If the conditional mutual information $I(\mathbf{Y}; \mathbf{Y}^{(s)}|\mathbf{X}_a) > 0$, we have $b_2 \geq b_1$ and $c_2 > c_1$, which means our approach strictly improves the bound for classification accuracy.*

We here use Fano’s Inequality to connect entropy to our prediction accuracy

Venn diagram for our defense



Our defense still work when adaptive attack know our defense

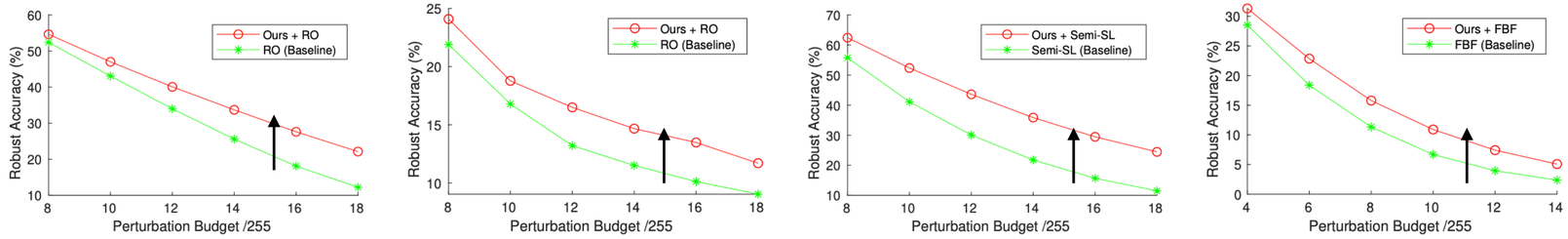


The state-of-the-art Defense

	CIFAR10	CIFAR100	SVHN	ImageNet
STOA Defense	60.2%	27.6%	53.6%	27.81%
Adding Ours	+7.5%	+5.5%	+11.8%	+3.1%

Experiment Result

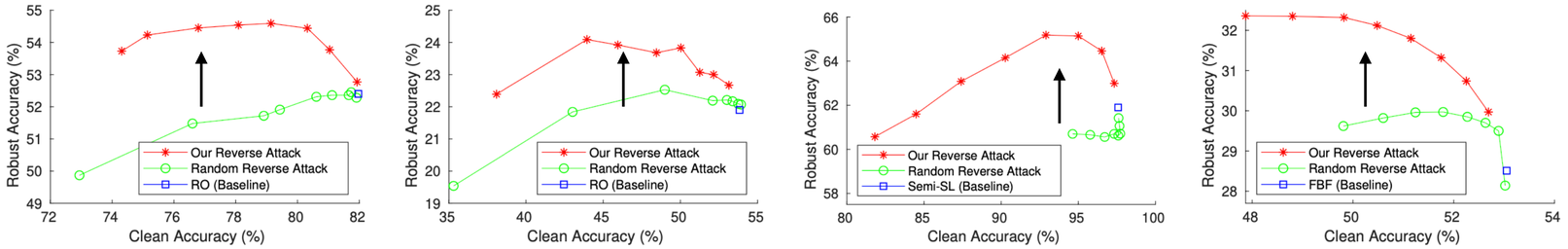
Our defense generalize to larger adversarial attack budget



(a) CIFAR-10 (up to 9.2% gain) (b) CIFAR-100 (up to 3.3% gain) (c) SVHN (up to 14.1% gain) (d) ImageNet (up to 4.4% gain)

Figure 6: The adversarial robust accuracy vs. perturbation budget curves on CIFAR-10, CIFAR-100, SVHN, and ImageNet, under the L_∞ norm. The red line is applying our inference algorithm to the baseline models [50, 61, 65]. Using our inference algorithm significantly improves the robustness.

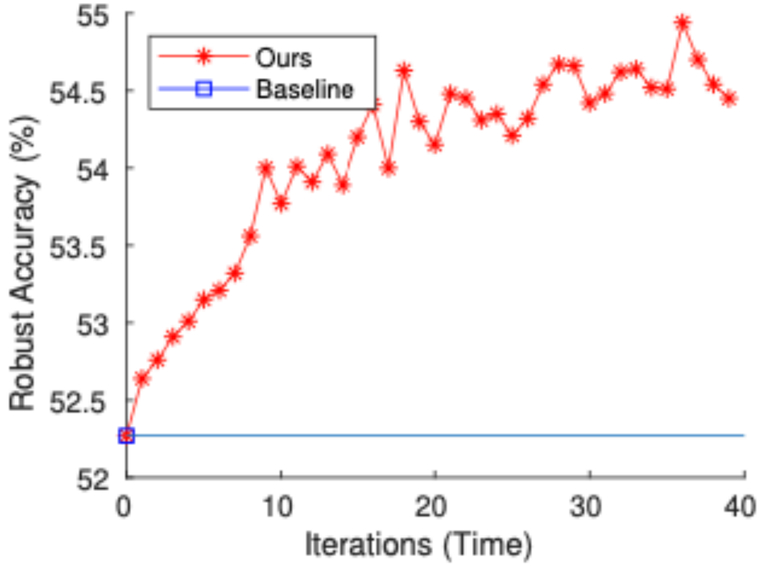
Our approach has better trade-off between robustness and clean accuracy



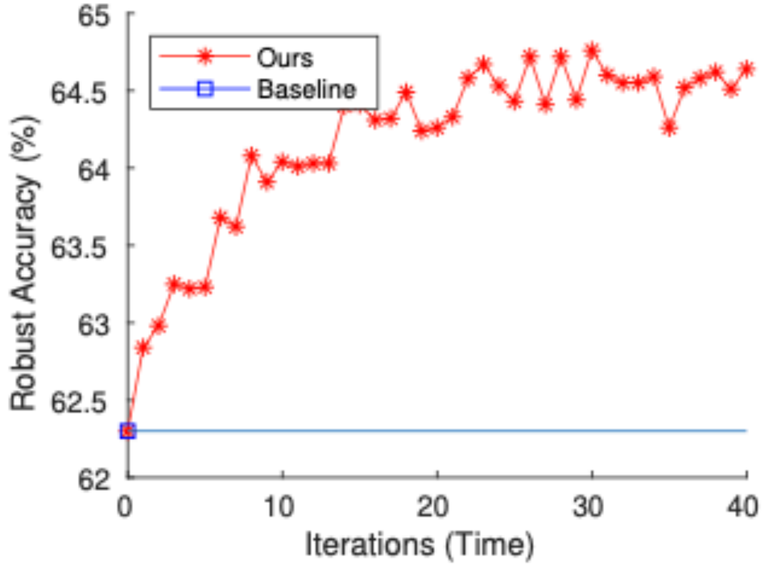
(a) CIFAR-10 (b) CIFAR-100 (c) SVHN (d) ImageNet

Figure 7: The trade-off between adversarial robust accuracy vs. clean accuracy on CIFAR-10, CIFAR-100, and SVHN under the L_∞ norm. We increase the noise budget ϵ_v from small to large, which causes the clean accuracy to drop from right to left. Our method produces a better reversal of the adversarial perturbation than just adding random noise to reverse it.

Steps for Reversal

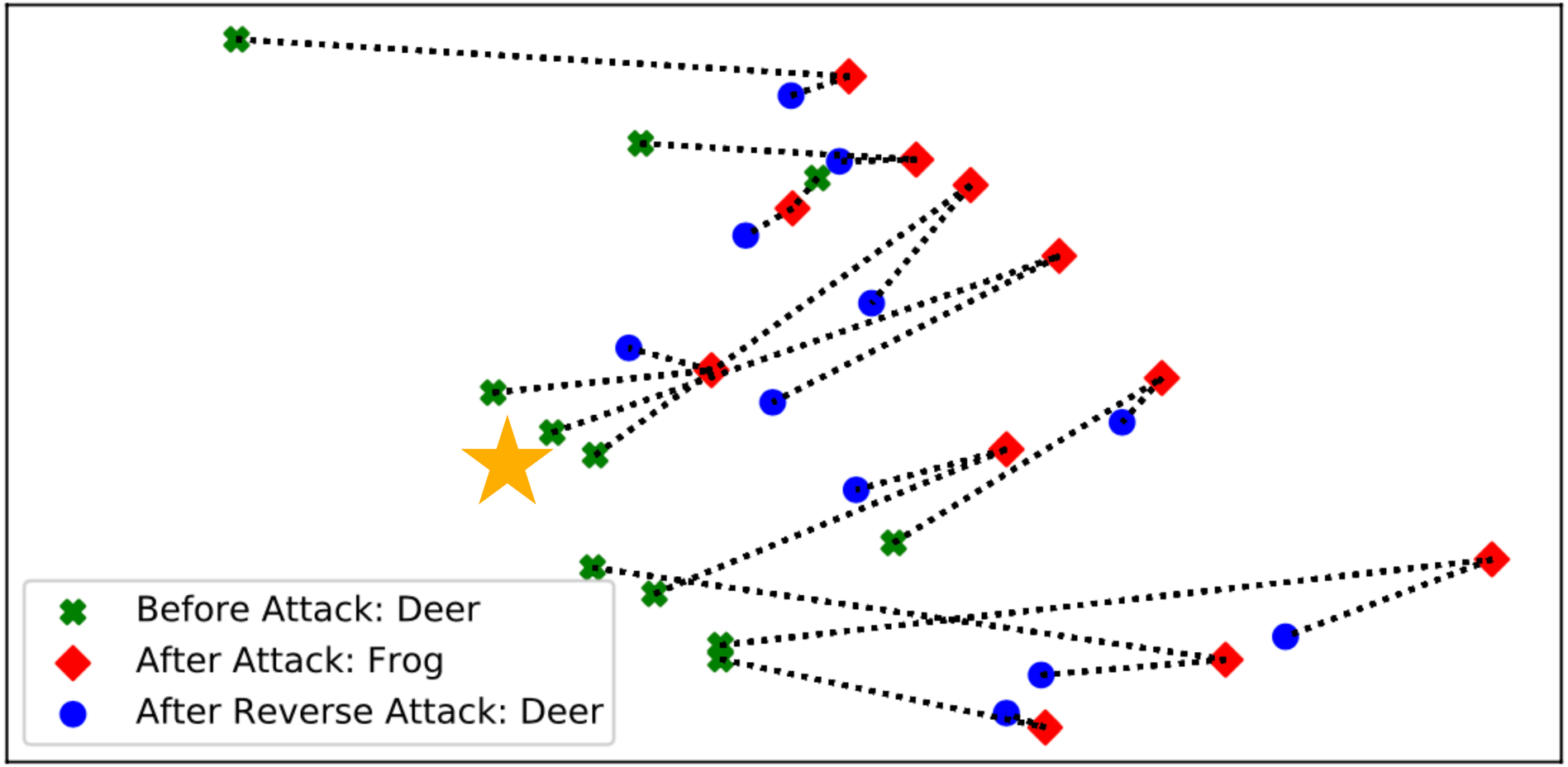


(a) RO [50]

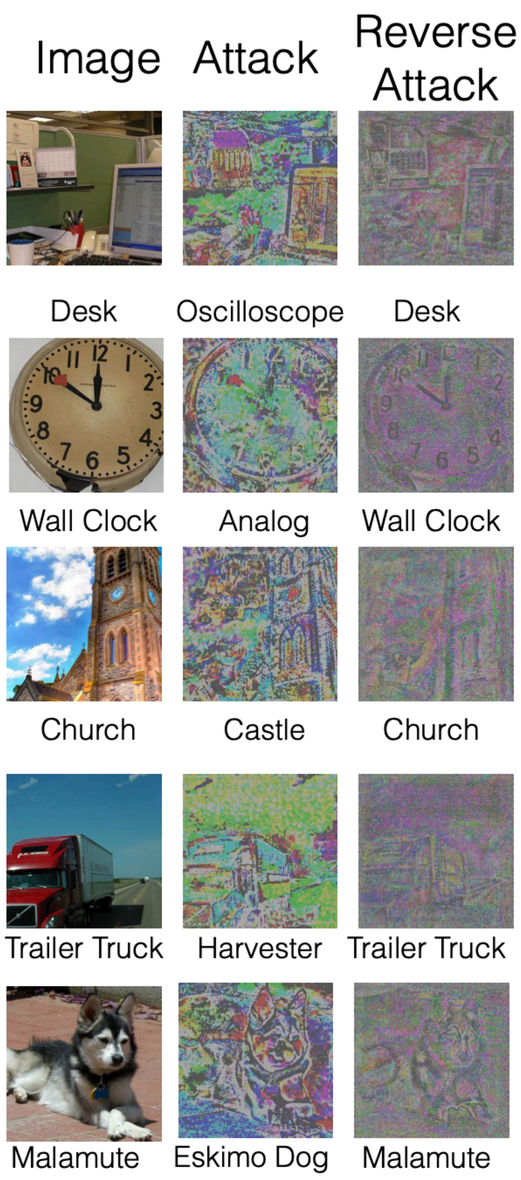


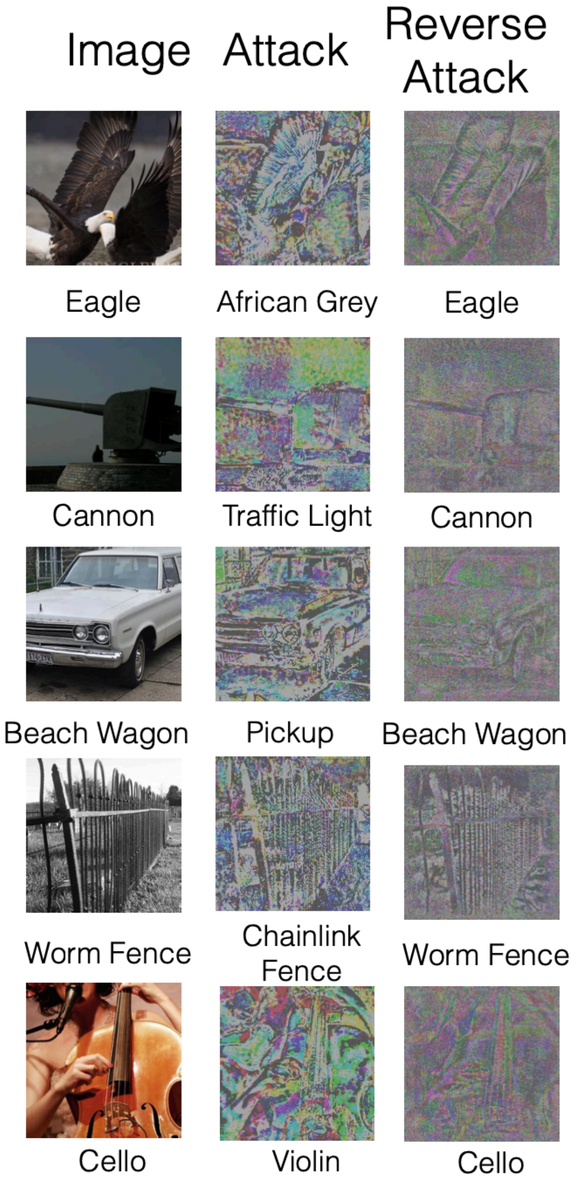
(b) Semi-SL [61]

Visualization for feature space before and after reversal



Visualization





Conclusion and future work

1. We propose an inference-time defense for adversarial attacks by restoring the intrinsic structure of the input.
2. Our method achieves the state-of-the-art defense on the auto-attack leaderboard
3. Our test time defense method is generic, which is compatible with existing training-time defense method, and can be used in other modalities such as sound, language, and beyond.

Code: https://github.com/cvlab-columbia/SelfSupDefense
Paper: https://arxiv.org/pdf/2103.14222.pdf