



Towards Robust Fine-grained Recognition by Maximal Separation of Discriminative Features

Asian Conference of Computer Vision 2020

Krishna Kanth Nakka and Mathieu Salzmann



1. Interpreting Adversarial Attacks:

We analyze and explain the decisions of fine-grained recognition networks by studying the image regions responsible for classification for both clean and adversarial examples.

2. Adversarial Defense:

We design an interpretable, attention-aware network for robust fine-grained recognition by constraining the latent space of discriminative regions.

Modules: Interpretable Fine-grained Network



ProtoPNet Architecture

Chaofan Chen et al. This Looks Like That: Deep Learning for Interpretable Image Recognition, NIPS 2019

Modules: Interpretable Fine-grained Network



Attention Pooling Architecture

For K- classes,

- K class specific filters
- 1 class agnostic filter
- Both attention maps are multiplied and spatially averaged to yield logits

Rohit Girdhar et al. Attentional Pooling for Action Recognition, NIPS 2017

• Discriminative regions of two different classes being too close in feature space

Input Normal Adversarial

Black-footed albatross Laysan albatross



Black-footed albatross



Black-footed albatross Laysan albatross





Rohit Girdhar et al. Attentional Pooling for Action Recognition, NIPS 2017

• Discriminative regions of two different classes being too close in feature space

Input

Normal Adversarial

Black-footed albatross Laysan albatross





Black-footed albatross Laysan albatross







• Discriminative regions of two different classes being too close in feature space



Chaofan Chen et al. This Looks Like That: Deep Learning for Interpretable Image Recognition, NIPS 2019

Use of non-discriminative regions for classification ٠



Input



Black-footed

prototype activation



Clean



Black-footed albatross



prototype source (pred label)

Adversarial



Least Auklet



Laysan albatross



prototype activation



prototype source (pred label)

Chaofan Chen et al. This Looks Like That: Deep Learning for Interpretable Image Recognition, NIPS 2019

Framework

We introduce an attention-based regularization mechanism

- Maximally separate the latent features of discriminative regions of different classes
- Minimize the contribution of the non-discriminative regions

Two regularization losses:

Attentional-cluster cost-Prototypes lie close to high attention regions of its own classAttentional-separation cost -Prototypes lie away from high attention regions of other classes

Architecture

Two prediction heads.

- 1. Spatial attention branch to obtain discriminative regions Ours-A
- 2. Feature regularization branch to maximally separate discriminative regions Ours-FR

Prototype Similarity maps are modulated with spatial attention branch to learn prototypes close to high attention regions.

Proposed Architecture



Attentional Cluster Loss:

The attentional-clustering loss pulls the high-attention regions in a sample close to the nearest prototype of its own class.

$$\mathcal{L}_{clst}^{att}(\mathbf{I}_i) = \sum_{t=1}^{N} a_i^t \min_{l:\mathbf{p}_l \in \mathbf{P}_{y_i}} \|\mathbf{x}_i^t - \mathbf{p}_l\|_2^2$$

$$m{a}_i^t$$
 - Attention weight at location t for image \mathbf{I}_i
 \mathbf{x}_i^t - feature vector at location t for image \mathbf{I}_i
 $\mathbf{P}_{y_i}^t$ - Set of prototypes belonging to class y_i

Discriminative Feature Separation

Attentional Separation Loss:

The attentional separation loss pushes the high-attention regions away from the nearest prototype of any other class.

$$\mathbf{L}_{sep}^{att}(\mathbf{I}_i) = -\sum_{t=1}^{N} a_i^t \min_{l: \mathbf{p}_l \notin \mathbf{P}_{y_i}} \|\mathbf{x}_i^t - \mathbf{p}_l\|_2^2$$

$$a_i^t$$
 - Attention weight at location t for image \mathbf{I}_i
 \mathbf{x}_i^t - feature vector at location t for image \mathbf{I}_i
 $\mathbf{P}_{y_i}^t$ - Prototypes belonging to class y_i

Discriminative Feature Separation

Combined regularization Loss:

We further push the non-discriminative regions away from informative prototypes by using attention from other images of Batch

$$\mathcal{L}_{reg}(\mathbf{I}_i) = \sum_{j=1}^B \sum_{t=1}^N \lambda_1 a_j^t \min_{l:\mathbf{p}_l \in \mathbf{P}_{y_i}} \|\mathbf{x}_i^t - \mathbf{p}_l\|_2^2 - \lambda_2 a_j^t \min_{l:\mathbf{p}_l \notin \mathbf{P}_{y_i}} \|\mathbf{x}_i^t - \mathbf{p}_l\|_2^2$$

Total Loss:

$$L(\mathbf{I}_i) = CE_{att}(\mathbf{I}_i) + CE_{reg}(\mathbf{I}_i) + L_{reg}(\mathbf{I}_i)$$

t-SNE Visualization of Learned Prototypes



Prediction: Blackfooted Albatross 10 (label: 0)

Clean

Prediction: Blackfooted Albatross (label: 0)

6

8

Experiments

Datasets – CUB200, Cars196 cropped images

Attacks - FGSM, BIM, PGD and MIM

Black Box Transfer attacks - BB-V (VGG-16) and BB-D (DenseNet)

Defense - Adversarial training with single step FGSM with random initialization*

Comparison of the Prototypes

ProtoPNet: multiple background prototypes and prototypes that focus on large regions



Comparison of the Prototypes

ProtoPNet: multiple background prototypes and prototypes that focus on large regions



Learned Prototypes

Prototypes learned by our attention-aware system



Ours: prototypes are finegrained and entire nondiscriminative regions is activated by single prototype

Learned Prototypes



Ours: prototypes are finegrained and entire nondiscriminative regions is activated by single prototype

Comparison of the activated image regions

ProtoPNet and Attentional Pooling



Results on CUB200 on Undefended Models

Base Network	$egin{array}{c} { m Attacks} \ { m (Steps,}\epsilon) \end{array}$	$\begin{array}{c} \text{Clean} \\ (0,0) \end{array}$	$\begin{array}{c} \mathrm{FGSM} \\ (1,2) \end{array}$	FGSM (1,8)	BIM (10,2)	BIM (10,8)	$_{(10,2)}^{\mathrm{PGD}}$	$\begin{array}{c} \mathrm{PGD} \\ (10,8) \end{array}$	MIM (10,2)	MIM (10,8)	BB-V (10,2)	BB-D (10,8)
VGG-16	AP [14] AP+ Triplet [57] AP+ PCL [35] Ours-A	78.0% 81.0% 80.0% 80.4%	36.5% 49.5% 41.0% 47.2%	31.0% 36.6% 33.1% 40.2%	27.7% 33.5% 32.9% 40.0%	14.6% 11.2% 13.6% 23.2%	23.5% 26.5% 23.5% 35.3%	11.7% 8.50% 9.6% 21.8%	30.2% 37.7% 35.3% 42.2%	16.7% 14.3% 17.1% 26.4%	9.6% 8.54% 10.6% 12.9%	60.4% 63.4% 65.8% 66.9%
	ProtoPNet [15] ProtoPNet+Ours	69.0% 73.2%	19.9% 49.9%	8.10% 42.2%	3.80% 42.5%	0.00% 35.3%	2.20% 38.4%	0.00% 30.1%	5.00% 42.9%	0.10% 37.5%	22.9% 15.4%	58.5% 59.7%
GG-19	AP [14] AP+ Triplet [57] AP+ PCL [35] Ours-A	75.7% 82.0% 76.9% 79.7%	20.4% 53.9% 20.3% 51.4%	$14.5\% \\ 38.2\% \\ 14.8\% \\ \mathbf{44.6\%} \\$	13.4% 35.0% 12.1% 42.3%	6.9% 12.4% 5.7% 26.5%	10.5% 27.7% 8.8% 36.8%	5.7% 9.40% 4.2% 26.3%	14.8% 39.4% 13.9% 45.0%	6.9% 15.3% 6.8% 42.6%	21.1% 17.40% 19.8% 29.8%	61.3% 64.9% 60.2% 68.2%
Ň	ProtoPNet [15] Ours-FR	73.8% 75.4%	22.9% 52.2%	11.1% 46.3%	3.2% 46.6%	0.0% 41.3%	1.2% 42.4%	0.0% 31.0%	3.6% 44.4%	0.0% 37.6%	21.0% 30.4%	58.0% 63.7%
ResNet-34	AP [14] AP+ Triplet [57] AP+ PCL [35] Ours-A	79.9% 78.6% 77.9% 79.0%	30.4% 25.6% 30.1% 32.3%	26.3% 18.7% 24.5% 27.0%	18.0% 11.4% 21.4% 24.8%	7.20% 2.9% 13.3% 20.5%	13.2% 7.1% 17.6% 22.5%	5.8% 1.8% 11.6% 19.8%	22.3% 14.7% 23.9% 26.2%	8.6% 3.8% 15.3% 22.0%	43.0% 42.11% 45.7% 48.6%	59.4% 58.4% 61.4% 63.2%
	ProtoPNet [15] Ours-FR	75.1% 76.3%	23.2% 30.7%	12.8% 22.0%	7.80% 19.3%	1.80% 13.6%	4.10% 14.2%	1.00% 13.0%	8.90% 19.1%	2.20% 13.8%	39.1% 46.0%	53.0% 60.0%

Table: Classification accuracy of different undefended networks with Linf basedattacks on CUB200

Results on Cars196 on Undefended Models

Base Nettwork	$egin{array}{c} { m Attacks} \ ({ m Steps},\epsilon) \end{array}$	$\begin{array}{c} { m Clean} \ (0,0) \end{array}$	$\begin{array}{c} {\rm FGSM} \\ (1,2) \end{array}$	FGSM (1,8)	BIM (10,2)	BIM (10,8)	$\begin{array}{c} \mathrm{PGD} \\ (10,2) \end{array}$	PGD (10,8)	MIM (10,2)	MIM (10,8)	BB-V (10,2)	BB-D (10,8)
GG-16	AP [14] AP+Triplet [57] AP+PCL [35] Ours-A	91.2% 91.1% 90.2% 88.5%	52.6% 54.3% 51.7% 58.7%	$\begin{array}{c} 40.2\%\\ \textbf{43.5\%}\\ 40.5\%\\ 40.2\%\end{array}$	37.4% 42.4% 39.3% 48.0%	10.5% 14.9% 14.1% 28.6%	28.8% 34.1% 31.8% 46.5%	6.93% 9.54% 9.44% 21.7%	41.7% 45.5% 42.5% 53.2%	12.9% 19.2% 17.5% 33.2%	12.5% 15.6% 16.7% 19.9%	82.5% 84.7% 83.9% 82.2%
>	ProtoPNet [15]	84.5%	31.2%	9.85%	4.78%	0.01%	2.23%	0.00%	6.5%	0.01%	27.8%	75.5%
	Ours-FR	83.8%	60.1%	52.0%	51.3%	41.0%	47.8%	32.9%	51.8%	43.9%	23.4%	75.1%
GG-19	AP	91.5%	50.1%	37.8%	33.4%	10.3%	23.83%	6.93%	37.9%	12.7%	20.7%	82.8%
	AP+Triplet [57]	91.0%	56.2%	45.1%	40.5%	13.0%	30.3%	8.70%	45.3%	16.7%	29.0%	85.0%
	AP+PCL [35]	91.3%	61.3%	49.9%	49.0%	19.7%	40.2%	14.1%	52.4%	23.4%	30.6%	85.7%
	Ours-A	88.7%	64.4%	54.8%	56.4%	36.7%	51.7%	33.4%	58.1%	41.0%	35.9%	82.5%
>	ProtoPNet [15]	85.6%	34.1%	20.8%	11.3%	1.11%	4.40%	0.5%	14.2%	1.39%	26.5%	75.5%
	Ours-FR	85.0%	62.4%	54.7%	54.5%	45.7%	51.2%	38.5%	54.3%	47.6%	36.1%	76.8%

Table: Classification accuracy of different undefended networks with Linf basedattacks on Cars196.

Results on CUB200 on Robust Models

$egin{array}{c} { m Attacks} \ { m (Steps,}\epsilon) \end{array}$	$_{(0,0)}^{ m Clean}$	FGSM $(1,2)$	FGSM $(1,8)$	$_{(10,2)}^{\rm BIM}$	BIM (10,8)	PGD (10,2)	$\begin{array}{c} \mathrm{PGD} \\ (10,8) \end{array}$	MIM (10,2)	MIM (10,8)	BB-V (10,2)	BB-D (10,8)
AP^{*} [14]	54.9%	44.9%	24.2%	41.9%	18.2%	41.2%	16.9%	41.9%	18.7%	54.6%	54.0%
$AP+PCL^{*}$ [35]	60.7%	50.5%	28.5%	47.1%	22.8%	46.7%	21.6%	47.2%	23.5%	59.5%	59.9%
Ours-A [*]	63.1%	5 6.1%	34.8%	51.7%	29.6%	50.8%	28.0%	52.0%	3 2.5%	66.3%	68.0%
ProtoPNet [*] [15]	60.1%	44.5%	26.9%	57.1%	10.9%	35.9%	10.3%	37.6%	13.5%	58.4%	59.1%
Ours-FR [*]	63.0%	5 3.3%	37.3%	49.4%	30.4%	5 48.1%	3 28.6%	49.7%	3 1.1%	61.1%	6 2.0%
AP^{*} [14]	58.0%	47.5%	29.1%	44.3%	25.6%	44.0%	24.34%	44.4%	26.2%	57.0%	57.3%
$AP+PCL^{*}$ [35]	61.8%	52.1%	30.9%	48.9%	24.7%	48.6%	23.3%	49.1%	25.4%	60.5%	60.9%
Ours-A [*]	68.2%	57.1%	36.5%	5 3.2%	30.4%	52.6%	2 9.2%	53.5%	3 1.2%	66.2%	6 66.9%
ProtoPNet [*] [15] Ours-FR [*]	55.1%	40.0%	28.9%	26.5%	11.3%	29.7%	9.60%	25.6%	10.2%	53.6%	53.9%
	64.4%	5 5.5%	37.4%	5 1.2%	30.6%	5 50.4%	3 28.7%	52.1%	3 2.3%	6 2.5%	6 3.2%
AP^{*} [14] $AP+PCL^{*}$ [35] Ours-A [*]	55.6% 54.5% 62.2%	47.8% 45.4% 5 4.2%	29.2% 26.9% 35.7%	$\begin{array}{c} 44.80\% \\ 42.3\% \\ 51.5\% \end{array}$	21.0% 18.2% 25.5%	$\begin{array}{c} 44.5\% \\ 41.9\% \\ 51.0\% \end{array}$	$19.4\% \\ 16.4\% \\ \mathbf{23.1\%}$	44.9% 42.4% 51.6%	21.9% 19.1% 2 6.6%	55.3% 54.0% 6 1.5%	55.2% 54.0% 61.9%
ProtoPNet [*] [15]	57.9%	46.5%	30.3%	41.1%	21.1%	40.3%	18.4%	41.5%	20.9%	56.9%	57.0%
Ours-FR [*]	57.6%	49.5%	32.3%	• 45.8%	23.2%	5 44.9%	5 19.9%	46.1%	2 4.6%	57.1%	57.0%
	Attacks $(Steps,\epsilon)$ AP* [14] AP+PCL* [35] Ours-A* ProtoPNet* [15] Ours-FR* AP* [14] AP+PCL* [35] Ours-FR* ProtoPNet* [15] Ours-A* ProtoPNet* [35] Ours-A* ProtoPNet* [15] Ours-FR*	Attacks Clean $(Steps,\epsilon)$ $(0,0)$ AP^* [14] 54.9% $AP+PCL^*$ [35] 60.7% $Ours-A^*$ 63.1% $ProtoPNet^*$ [15] 60.1% $Ours-FR^*$ 63.0% AP^* [14] 58.0% $AP+PCL^*$ [35] 58.0% $Ours-A^*$ 64.4% $ProtoPNet^*$ [15] 55.6% AP^* [14] 54.5% $Ours-FR^*$ 64.4% AP^* [14] 54.5% $Ours-A^*$ 54.5% $Ours-A^*$ 57.6% $ProtoPNet^*$ [15] 57.9% $Ours-FR^*$ 57.6%	AttacksCleanFGSM $(3 \text{teps}, \epsilon)$ $(0,0)$ $(1,2)$ AP^* $[14]$ 54.9% 44.9% $AP+PCL^*$ $[35]$ 60.7% 50.5% $Ours-A^*$ 63.1% 56.1% $ProtoPNet^*$ $[15]$ 60.1% 44.5% $Ours-FR^*$ 63.0% 53.3% AP^* $[14]$ 58.0% 47.5% $AP+PCL^*$ $[35]$ 61.8% 52.1% $Ours-A^*$ 64.4% 55.5% AP^* $[14]$ 55.6% 47.8% $AP+PCL^*$ $[35]$ 55.6% 47.8% AP^* $[14]$ 55.6% 45.4% $Ours-A^*$ 55.6% 45.4% $Ours-A^*$ 57.9% 46.5% $ProtoPNet^*$ $[15]$ 57.9% 46.5% $Ours-FR^*$ 57.6% 49.5%	AttacksCleanFGSMFGSM $(Steps,\epsilon)$ $(0,0)$ $(1,2)$ $(1,8)$ AP^* $[14]$ 54.9% 44.9% 24.2% $AP+PCL^*$ $[35]$ 60.7% 50.5% 28.5% $Ours-A^*$ 63.1% 56.1% 34.8% $ProtoPNet^*$ $[15]$ 60.1% 44.5% 26.9% $Ours-FR^*$ 63.0% 53.3% 37.3% AP^* $[14]$ 58.0% 47.5% 29.1% $AP+PCL^*$ $[35]$ 61.8% 52.1% 30.9% $Ours-A^*$ 68.2% 57.1% 36.5% $ProtoPNet^*$ $[15]$ 55.6% 47.8% 29.2% AP^* $[14]$ 55.6% 47.8% 29.2% AP^* $[14]$ 55.6% 47.8% 29.2% $AP+PCL^*$ $[35]$ 54.5% 45.4% 26.9% $Ours-A^*$ 62.2% 54.2% 35.7% $ProtoPNet^*$ $[15]$ 57.9% 46.5% 30.3% $Ours-FR^*$ 57.6% 49.5% 32.3%	Attacks (Steps, ϵ)Clean (0,0)FGSM (1,2)FGSM (1,8)BIM (10,2)AP*[14] AP+PCL*54.9% 60.7%44.9% 50.5%24.2% 28.5%41.9% 47.1% 63.1%Ours-A*60.7% 63.1%50.5% 56.1%28.5% 34.8%51.7% 51.7%ProtoPNet*[15] 60.1%44.5% 44.5%26.9% 26.9%57.1% 63.0%Ours-FR*63.0% 63.0%53.3% 53.3%37.3% 49.4%AP*[14] AP+PCL*58.0% 61.8%47.5% 52.1%29.1% 30.9% 48.9% 68.2%Ours-A*68.2% 64.4%55.5% 57.1%36.5% 53.2%AP*[14] AP+PCL*55.6% 54.5%47.8% 29.2% 44.80% 62.2%AP*[14] S5.6%47.8% 45.4%29.2% 44.80% 42.3% 62.2%ProtoPNet*[15] 57.9%46.5% 30.3%30.3% 41.1% 45.8%ProtoPNet*[15] 57.9%57.6% 49.5%32.3% 45.8%	Attacks (Steps, ϵ)Clean (0,0)FGSM (1,2)BIM (10,8)BIM (10,2)AP*[14] AP+PCL*54.9% (60.7%44.9% 50.5%24.2% 28.5%41.9% 41.9%18.2% 22.8% 22.8% 63.1%Ours-A*63.1% 63.1%56.1% 50.5%34.8% 51.7%29.6% 29.6%ProtoPNet*[15] 60.1%60.7% 53.3%26.9% 37.3%57.1% 49.4%10.9% 29.6%AP*[14] AP+PCL*58.0% 61.8%47.5% 52.1%29.1% 30.9%44.3% 48.9%25.6% 24.7%AP*[14] AP+PCL*55.1% 64.4%40.0% 55.5%28.9% 37.4%26.5% 51.2%11.3% 30.6%AP*[14] AP+PCL*55.6% 64.4%47.8% 55.5%29.2% 37.4%44.80% 51.2%21.0% 30.6%AP*[14] AP+PCL*55.6% 45.4%45.4% 26.9%26.5% 42.3%18.2% 18.2%ProtoPNet*[15] 54.5%45.4% 45.4%26.9% 26.9%42.3% 45.8%23.2%ProtoPNet*[15] 57.9%46.5% 45.4%30.3% 26.9%41.1% 21.1% 21.1%ProtoPNet*[15] 57.9%46.5% 49.5%30.3% 32.3%45.8% 45.8%23.2%	Attacks (Steps, ϵ)Clean (0,0)FGSM (1,2)BIM (1,8)BIM (10,2)PGD (10,8)AP* AP+PCL* Ours-A*54.9% 60.7%44.9% 50.5%24.2% 28.5%41.9% 47.1%18.2% 22.8%41.2% 46.7% 29.6%Ours-A*63.1% 63.1%56.1% 50.5%34.8% 51.7%51.7% 29.6%29.6% 50.8%ProtoPNet* Ours-FR*[15] 60.1%44.5% 44.5%26.9% 26.9%57.1% 57.1%10.9% 35.9% 30.4%AP* AP+PCL* Ours-A*[35] 61.8%52.1% 52.1%30.9% 30.9%48.9% 48.9%24.7% 48.6% 24.7%AP* AP+PCL* Ours-FR*[35] 64.4%55.5% 57.1%36.5% 53.2%30.4% 51.2%50.4% 30.6%AP* ProtoPNet* AP+PCL* AP+PCH* AP+PCH* AP+PCH* AP+PCH* A	Attacks (Steps, ϵ)Clean (0,0)FGSM (1,2)FGSM (1,8)BIM (10,2)PGD (10,8)PGD (10,2)PGD (10,8)AP* AP+PCL* Ours-A*54.9% 60.7%44.9% 50.5%24.2% 24.2%41.9% 41.9%18.2% 41.2%41.2% 41.9%16.9% 22.8%ProtoPNet* Ours-FR*[15] 60.1%60.1% 44.5%26.9% 26.9%57.1% 50.8%10.9% 29.6%50.8% 28.0%ProtoPNet* AP+PCL* AP+PCL* Ours-A*[15] 60.1%44.5% 44.5%26.9% 29.1%57.1% 44.3%10.9% 25.6%30.4% 48.1%24.34% 28.6%AP* Ours-A*[15] 61.8%52.1% 52.1%30.9% 30.9%48.9% 24.7%24.7% 48.6%23.3% 23.3%ProtoPNet* Ours-FR*[15] 55.1%40.0% 40.0%28.9% 26.5%26.5% 11.3%11.3% 29.7%9.60% 9.60%AP* AP+PCL* AP+PCL* Ours-A*[15] 55.6%47.8% 45.4%29.2% 44.80%44.5% 19.4%19.4% 45.5%AP* ProtoPNet* Durs-A*[15] 57.9%46.5% 46.5%30.3% 30.3%41.1% 41.1%21.1% 40.3%18.4% 40.3%ProtoPNet* Durs-FR*[15] 57.9%57.6% 49.5%30.3% 30.3%41.1% 41.1%21.1% 40.3%18.4% 40.3%	Attacks (Steps, ϵ)Clean (0,0)FGSM (1,2)FGSM (1,8)BIM (10,2)PGD (10,8)PGD (10,2)PGD (10,8)MIM (10,2)AP* AP+PCL* Ours-A*54.9% 60.7%44.9% 50.5% 28.5%24.2% 41.9%41.9% 18.2%41.2% 41.2%16.9% 41.9%41.9% 47.2%Ours-A*63.1% 60.1%56.1% 34.8%34.8% 51.7%29.6% 50.8%50.8% 28.0%28.0% 52.0%ProtoPNet* Ours-FR*[15] 60.1%60.1% 44.5%26.9% 26.9%57.1% 50.8%10.9% 35.9%37.6% 49.4%AP* Ours-FR*[14] 61.8%52.1% 52.1%30.9% 48.9%24.7% 24.7%48.6% 23.3%23.3% 49.1%Ours-FR*64.4% 55.5%37.4% 51.2%30.4% 30.6%50.4% 28.7%52.1% 52.1%AP* ProtoPNet* (15]55.6% 54.5%47.8% 29.2%29.2% 44.80%44.5% 21.0%19.4% 44.9% 44.9%AP* ProtoPNet* (14] AP+PCL* (35]55.6% 54.5%47.8% 29.2%21.0% 44.80%44.5% 21.0%19.4% 44.9% 44.9%AP* ProtoPNet* (15]57.9% 54.5%46.5% 30.3%30.3% 41.1%21.1% 40.3%44.9% 44.9% 44.9%AP* ProtoPNet* (15]57.9% 57.9%46.5% 30.3%30.3% 41.1%21.1% 21.1%40.3% 40.3%18.4% 41.5%ProtoPNet* ProtoPNet* (15]57.9% 57.9%46.5% 30.3%30.3% 41.1%21.1% 21.1%40.3% 40.3%18.4	Attacks (Steps, ϵ)Clean (0,0)FGSM (1,2)FGSM (1,8)BIM (10,2)PGD (10,8)PGD (10,2)PGD (10,8)MIM (10,2)MIM (10,2)AP* AP+PCL* Ours-A*54.9% (60.7%44.9% 50.5%24.2% 28.5%41.9% 47.1%18.2% 22.8%41.2% 46.7%16.9% 21.6%41.9% 47.2%18.7% 23.5%Ours-A*60.7% 63.1%50.5% 50.5%28.5% 28.5%47.1% 22.8%22.8% 46.7%21.6% 21.6%47.2% 23.5%ProtoPNet* AP+PCL* AP+PCL* AP+PCL* AP+PCL* AP+PCL* B3544.5% 26.1%26.9% 29.1%57.1% 44.3% 25.6%10.9% 35.9%37.6% 24.3%33.6% 48.1%24.3% 24.3%44.4% 26.2%AP* AP+PCL* B3655.1% 64.4%40.0% 28.9%28.9% 24.7%24.3% 48.6%23.3% 23.3%49.1% 25.4%ProtoPNet* AP+PCL* AP+PCL* AP+PCL* B4+PCL* AP+PCL* B4+PCL* B3555.6% 56.6%47.8% 29.2% 44.80%21.0% 21.0%44.5% 44.5%19.4% 21.9%44.9% 21.9%AP* AP+PCL* B4+PCL*	Attacks (Steps, ϵ)Clean (0,0)FGSM (1,2)FGSM (1,8)BIM (10,2)PGD (10,8)PGD (10,2)MIM (10,2)MIM (10,2)BB-V (10,8)AP* AP+PCL* Ours-A*54.9% 60.7%44.9% 50.5%24.2% 28.5%41.9% 41.9%18.2% 41.2%41.9% 41.2%16.9% 41.9%41.9% 41.9%18.7% 54.6%Ours-A*60.7% 63.1%50.5% 56.1%28.5% 47.1%47.1% 22.8%22.8% 46.7%41.2% 21.6%16.9% 47.2%41.9% 23.5%18.7% 59.5%ProtoPNet* AP+PCL* AP+PCL* AP+PCL* B16156.1% 61.8%52.1% 29.1%29.6% 50.8%50.8% 28.0%28.0% 52.0%52.0% 37.6%32.5% 66.3%AP* AP+PCL* B1658.0% 61.8%47.5% 29.1%29.6% 44.3%50.8% 28.0%24.3% 48.1%44.4% 26.2%26.2% 57.0%AP* ProtoPNet* B1555.1% 61.8%40.0% 29.1%24.7% 48.6%48.6% 23.3%23.3% 49.1%25.4% 60.5%Ours-A*64.4% 55.5%37.4% 51.2%30.4% 30.6%52.6% 29.2%53.5% 53.5%31.2% 66.2%AP* ProtoPNet* B14 AP+PCL* AP+PCL* B1555.6% 54.5%47.8% 29.2%44.80% 21.0%21.0% 44.5%44.9% 21.9%21.9% 54.6%AP* AP+PCL* B14 AP+PCL* AP+PCL* B1555.6% 54.5%47.8% 29.2%29.2% 44.80%44.5% 20.0%19.4% 44.5%44.9% 21.9%21.9% 54.6%AP*

Table: Classification accuracy of different robust networks with Linf based attacks on CUB200.

Results on Cars196 on Robust Models

Base Network	$\begin{array}{l} \text{Attacks} \\ \text{Steps}, \epsilon \end{array} $	Clean	$\begin{array}{c} \mathrm{FGSM} \\ (1,2) \end{array}$	FGSM (1,8)	$_{(10,2)}^{\rm BIM}$	BIM (10,8)	$\begin{array}{c} \mathrm{PGD} \\ (10,2) \end{array}$	$\begin{array}{c} \mathrm{PGD} \\ (10,8) \end{array}$	MIM (10,2)	MIM (10,8)	BB-V (10,8)	BB-D (10,8)
3G-16	AP* [6]	86.2%	81.1%	63.6%	78.9%	53.8%	78.7%	50.8%	78.7%	55.1%	85.1%	85.9%
	AP+PCL* [7]	87.4%	80.5%	59.4%	77.6%	48.5%	77.2%	44.9%	77.9%	50.2%	86.0%	87.1%
	Ours-A *	84.8%	79.8%	63.3%	77.0%	54.6%	76.6%	51.1%	77.1%	55.8%	84.5%	85.6%
λ	ProtoPNet [*] [3]	64.4%	53.7%	31.9%	48.9%	16.5%	48.2%	13.4%	49.2%	18.2%	63.8%	64.2%
Λ	Ours-FR [*]	83.7%	76.37%	62.8%	73.5%	55.0%	72.6%	51.9%	73.8%	55.4%	80.8%	82.0%
G-19	AP* [6]	88.2%	82.4%	63.4%	79.9%	54.2%	796%	50.7%	80.0%	55.7%	86.9%	88.0%
	AP+PCL* [7]	88.2%	82.7%	64.6%	80.2%	57.4%	79.6%	54.3%	80.3%	58.5%	87.2%	88.1%
	Ours-A *	87.3%	80.29%	67.1%	78.4%	60.15%	78.2%	58.2%	78.6%	61.3%	86.5%	87.3%
2V	ProtoPNet [*] [3]	30.0%	19.9%	15.7%	15.0%	16.3%	9.1%	3.00%	3.32%	2.28%	29.4%	29.7%
	Ours-FR [*]	84.6%	79.6%	66.9%	77.7%	58.6%	76.5%	55.6%	77.8%	59.1%	83.7%	84.5%

Table: Classification accuracy of different robust networks with Linf based attacks onCars196.

Ablation Study

etwork .	Att-clustering A loss	Att-separation loss	$\begin{array}{c} { m Clean} \\ (0,0) \end{array}$	$\begin{array}{c} \mathrm{PGD} \\ (10,8) \end{array}$	Network	Att-clustering loss	$\begin{array}{c} \text{Att-separation} \\ \text{loss} \end{array}$	$\stackrel{ m Clean}{ m (0,0)}$	\mathbf{P} (1)
P [14]	-	-	78.0%	11.7%	ProtoPNet [15]	-	-	69.0%	0.
)urs-A	- - - -	- - - -	$78.7\% \\ 79.6\% \\ 80.0\% \\ 80.4\%$	$14.07\% \\ 0.0\% \\ 19.3\% \\ 21.8\%$	Ours-FR	- - - -	- - -	$75.7\% \\ 69.8\% \\ 73.7\% \\ 73.2\%$	$13. \\ 0. \\ 18 \\ 30$

Table: Contribution of each proposed feature regularization module in classification accuracy of undefended VGG-16 network

Gradient Obfuscation Study



Table: Performance of VGG-16 with our proposed approach under different perturbation strengths.

Adversarial Detection



Table: ROC curves for adversarial sample detection on robust VGG-16 with PGD attack Ours-A* and Ours-FR* performs better than baselines

- Compute minimum Mahalanobis distance from pretrained class conditional distributions at each layer
- Train a logistic detector on 20% samples and evaluated on rest 80% of adversarial successful cum correctly classified test data

Conclusion

- We have performed the first study of adversarial attacks for fine-grained recognition.
- Our analysis has highlighted the key factor for the success of adversarial attacks in this context.

• Designed an attention and prototype-based framework that explicitly encourages the prototypes to focus on the discriminative image regions

Thank you!