Feature Separation and Recalibration for Adversarial Robustness

Woo Jae Kim, Yoonki Cho, Junsik Jung, Sung-Eui Yoon

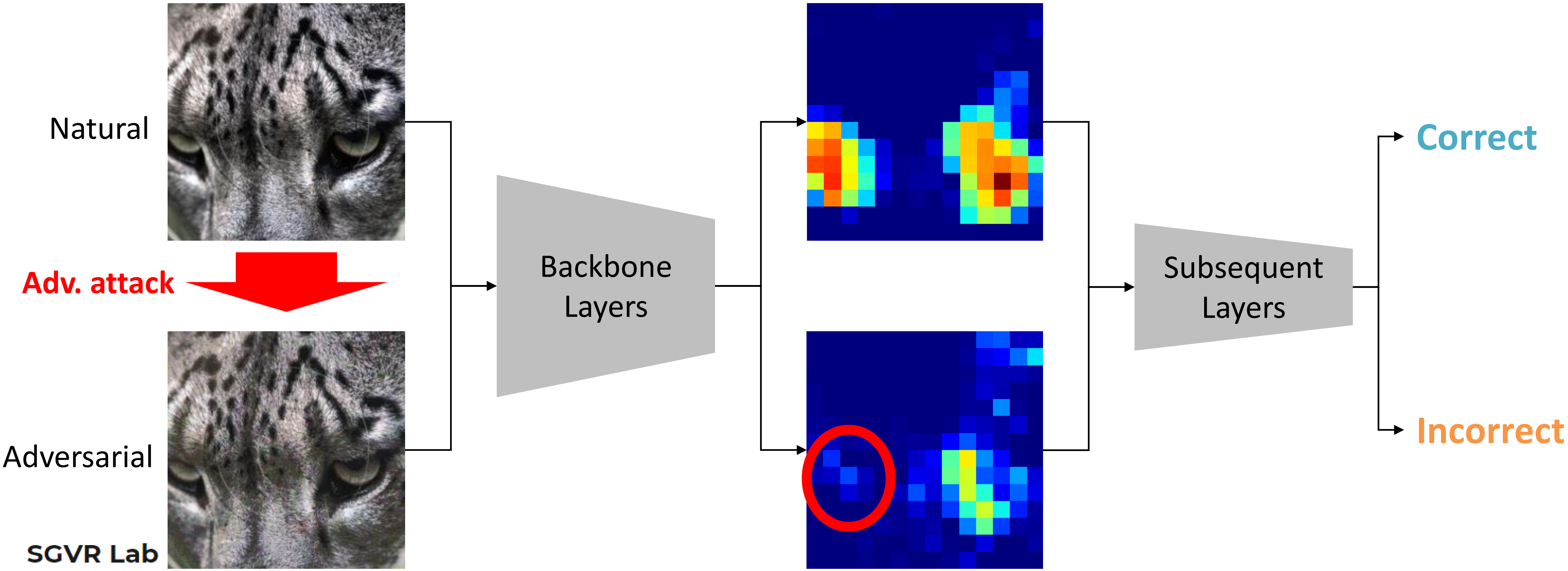
TUE-PM-389

CVPR 2023 (Highlights)

Preview

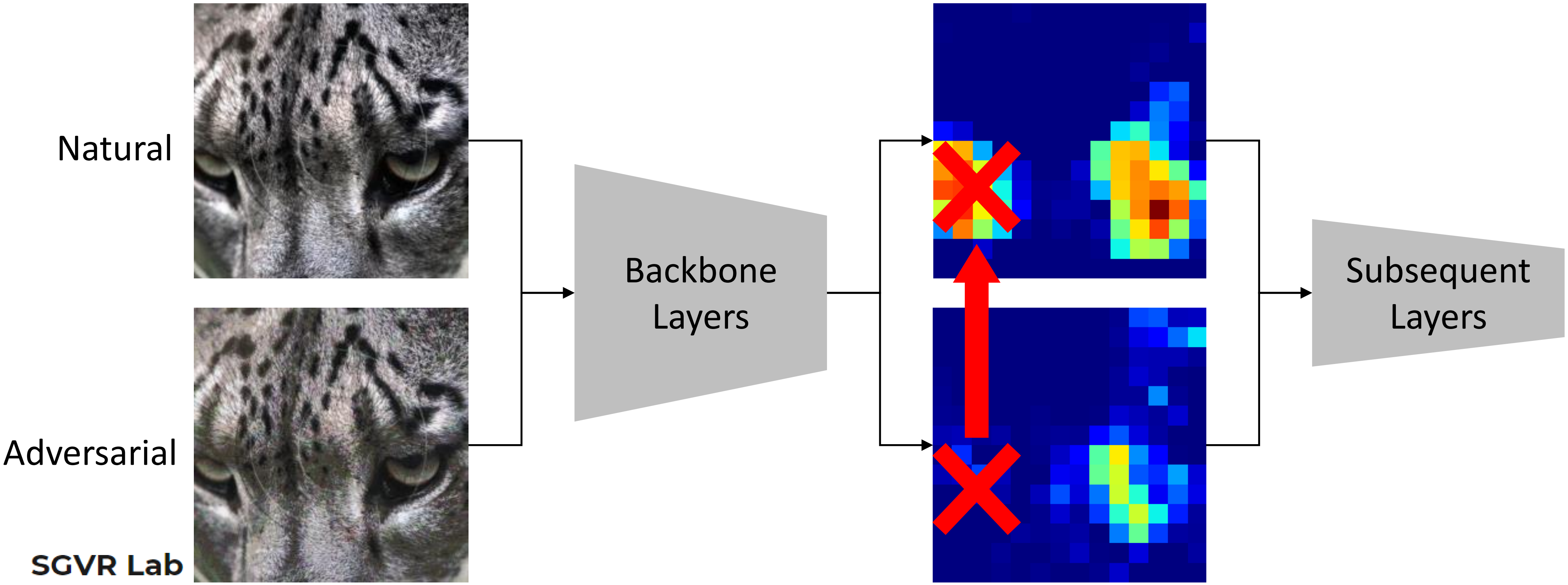
Feature Activation Disruption upon Adversarial Attack

- Feature-level disruptions lead to model mispredictions



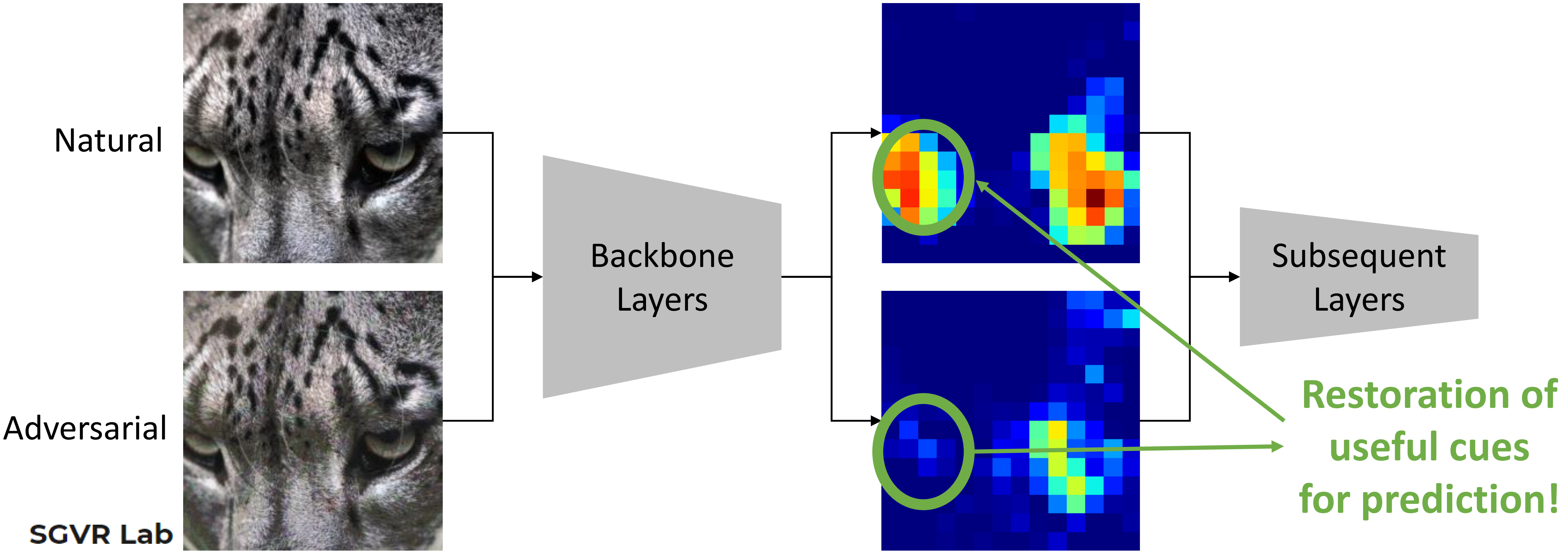
Limitations of Conventional Defense

- Conventional defense methods ***suppressed*** or ***deactivated*** disrupted activations
- This approach can lead to ***loss of potentially discriminative cues***



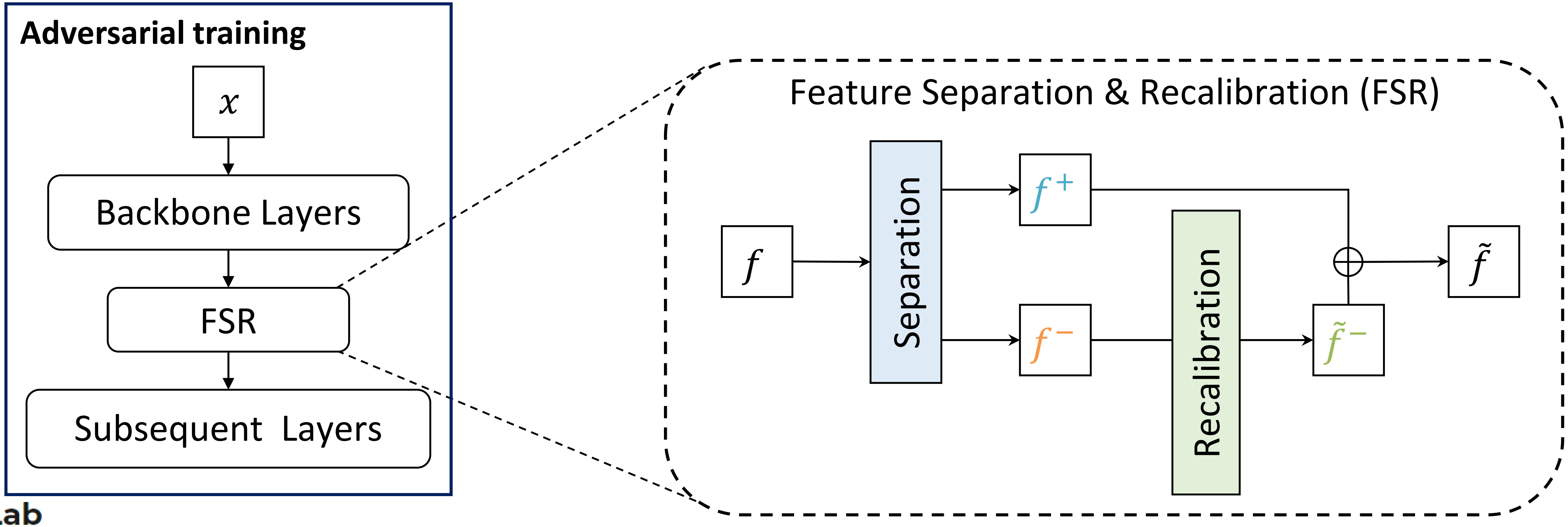
Proposed Approach

- Instead, we propose to ***restore useful cues*** from these disrupted activations
- These additional useful cues ***enrich*** model's ability to make ***correct predictions***



Feature Separation and Recalibration (FSR)

- Robust feature f^+: Useful cues
- Non-robust feature f^-: Malicious cues responsible for mispredictions
- Recalibrated feature $\tilde{f}^-$: Restored useful cues

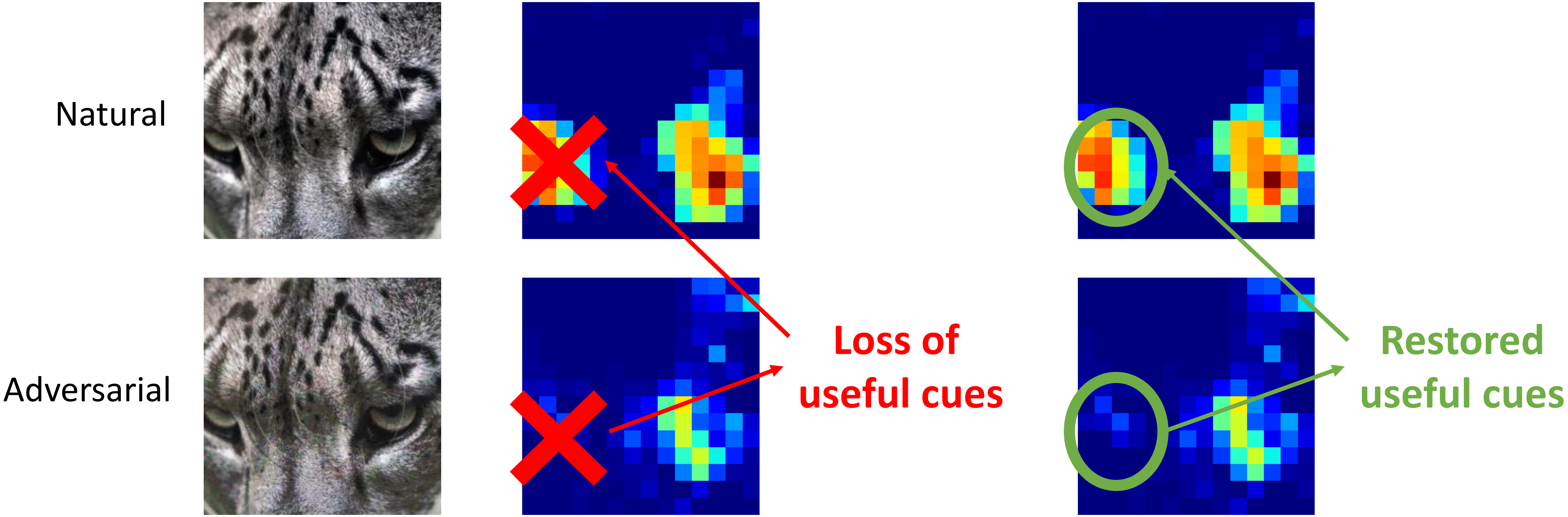


Proposed Approach

Feature Separation and Recalibration

Feature Activation Disruption upon Adversarial Attack

- Goal: Restore useful cues for correct predictions from disrupted activations
- These restored cues will provide richer information for making correct predictions

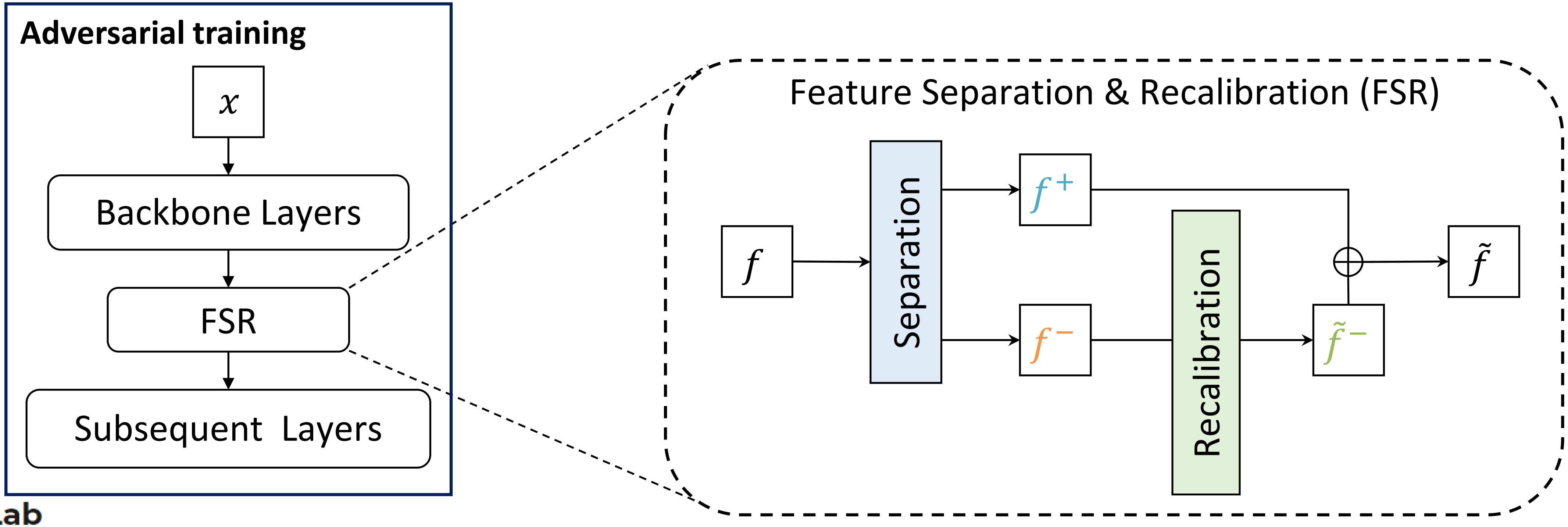


Conventional approach

Our approach

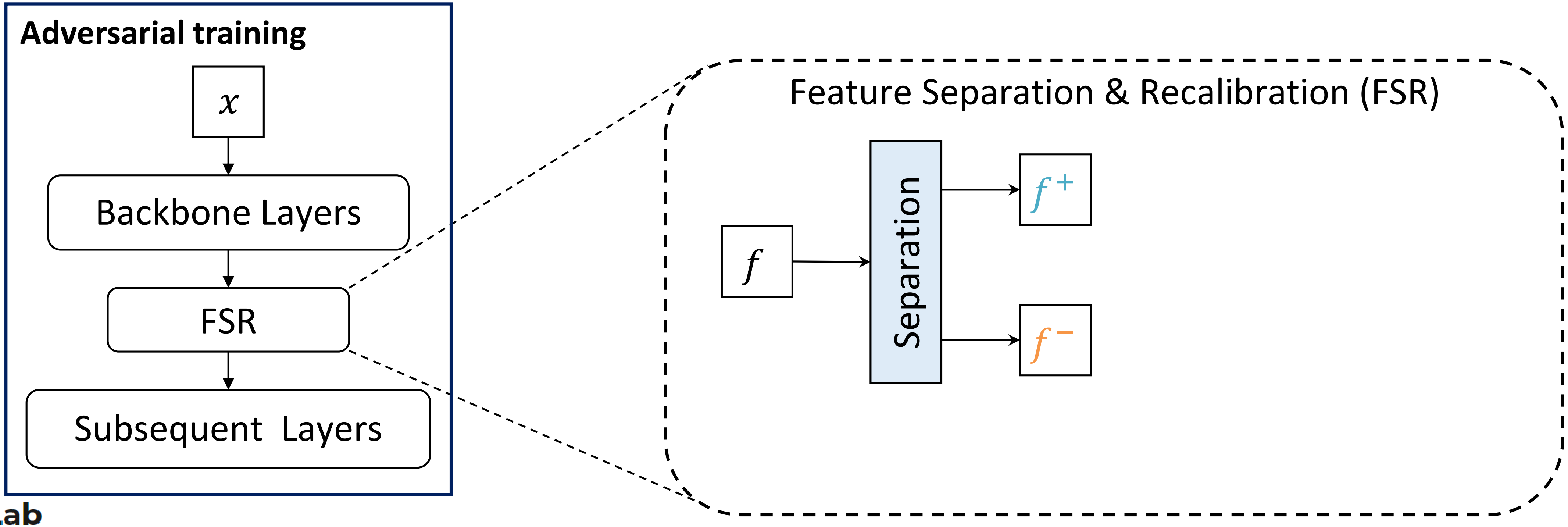
Feature Separation and Recalibration (FSR)

- Module inserted to ***any CNN model***
- Trained with ***any adversarial training*** technique in an ***end-to-end*** manner
- Recalibrates disrupted feature activations to restore useful cues for predictions



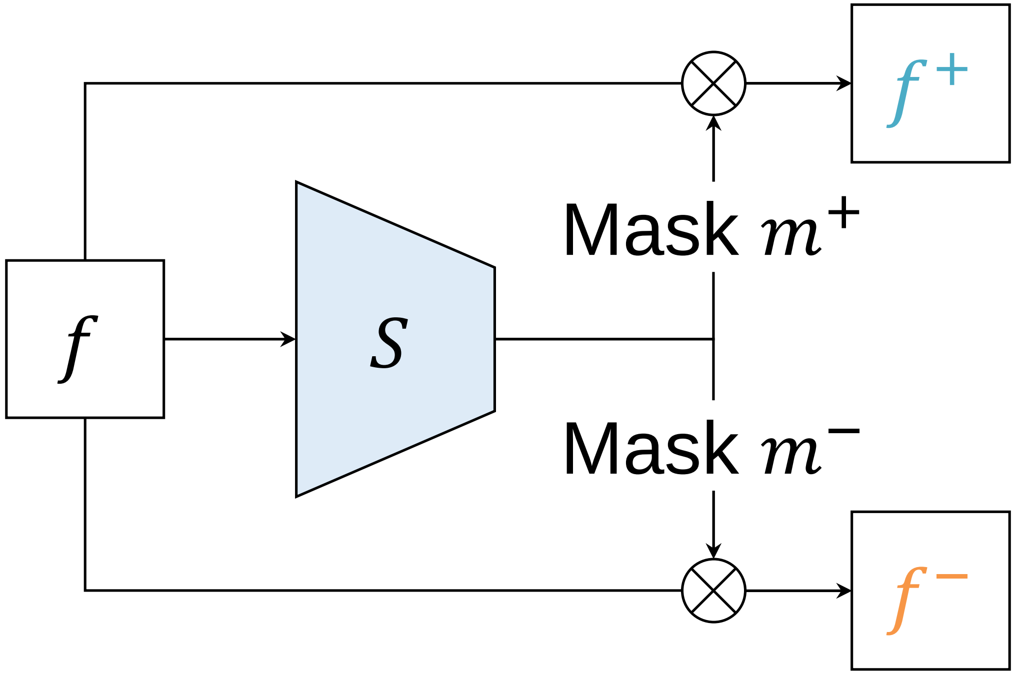
Feature Separation

- Separation: Separate feature f into robust feature f^+ and non-robust feature f^-
- Robust f^+: Activations that provide useful cues
- Non-robust f^-: Activations that are responsible for model mispredictions



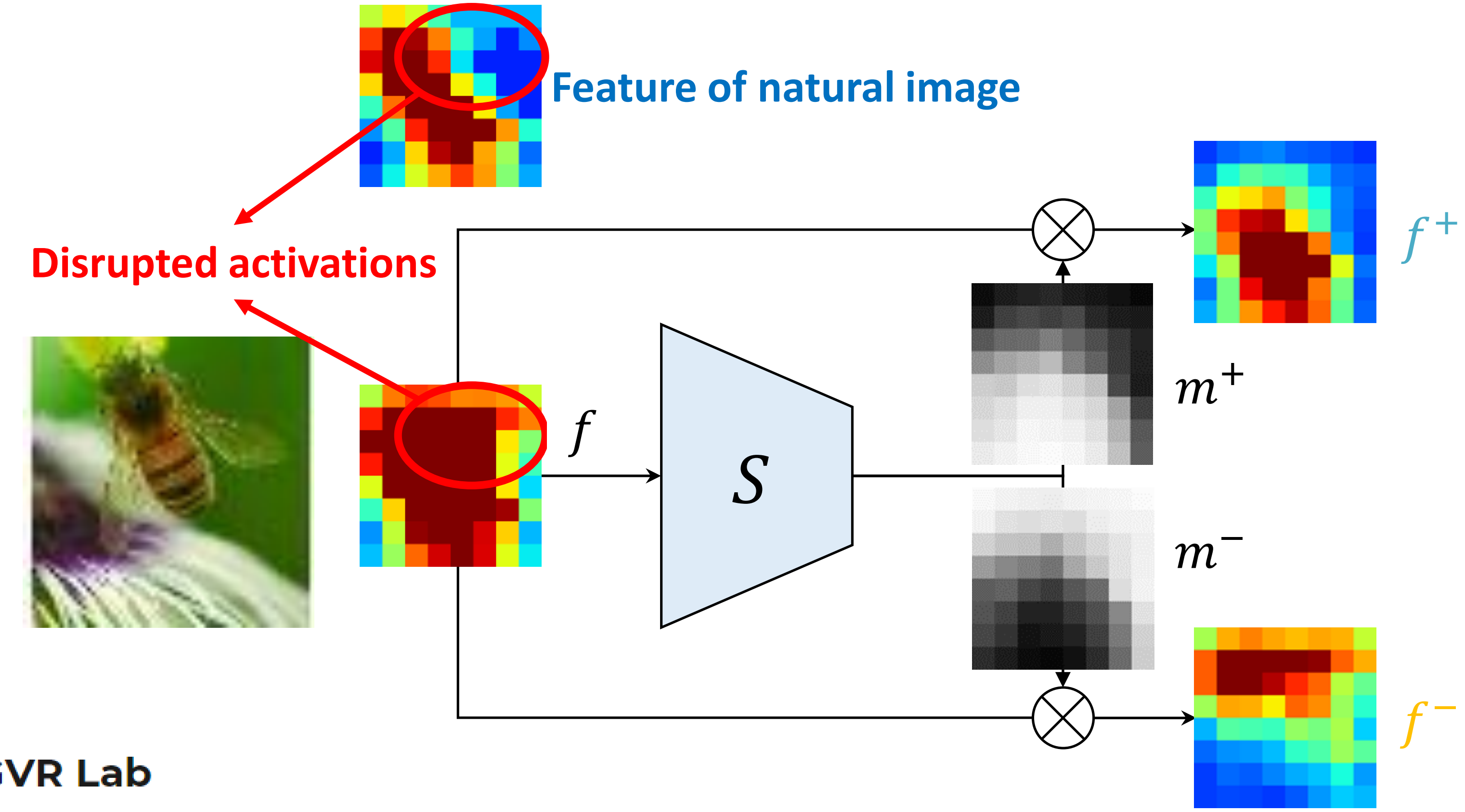
Feature Separation

- Separation Net S learns the robustness of each activation of input feature f
- We activation-wise separate the feature based on the robustness



Feature Separation

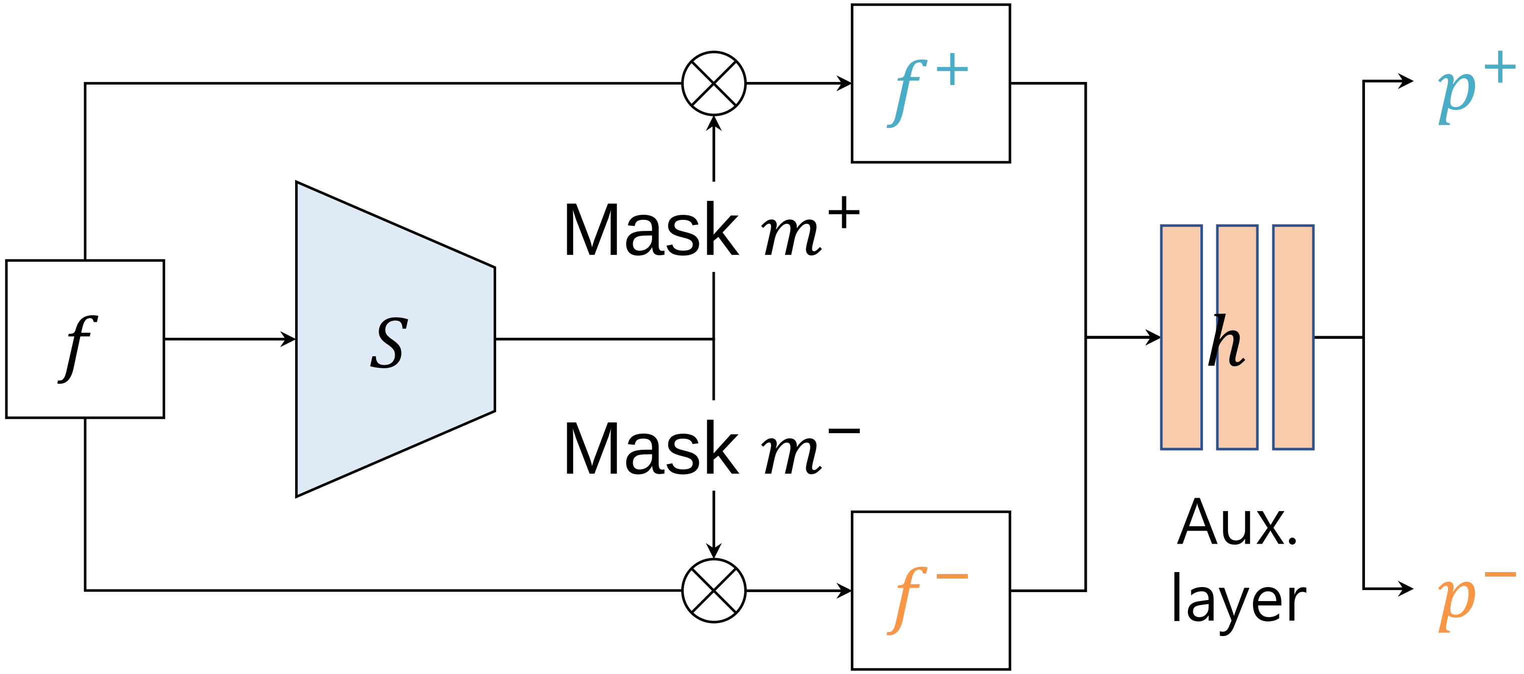
- Positive mask emphasizes activations relevant to correct predictions
- Negative mask emphasizes activations relevant to mispredictions



Feature Separation

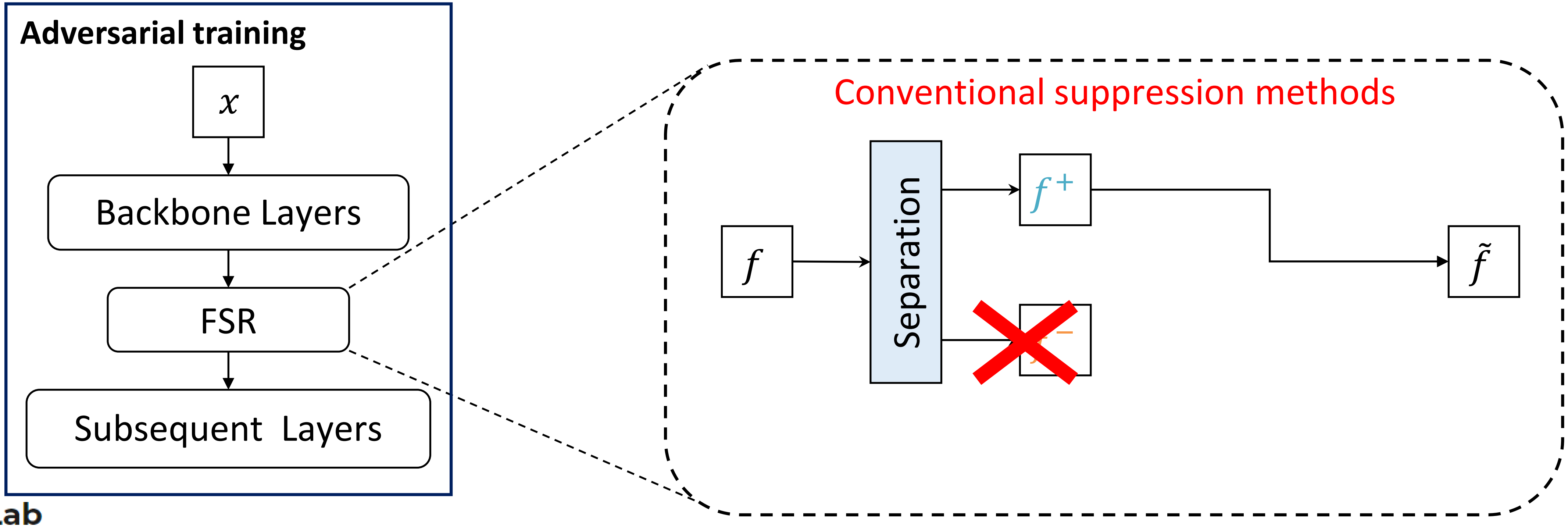
- Guide the Sep. Net S to learn robustness based on relevance to correct prediction

$$\mathcal{L}_{sep} = \underbrace{\mathcal{H}}_{\text{Cross-entropy loss}}(\underbrace{p^+}_{\text{Pred. logit}}, \underbrace{y}_{\text{GT label}}) + \mathcal{H}(\underbrace{p^-}_{\text{Pred. logit}}, \underbrace{y'}_{\text{Wrong label}})$$



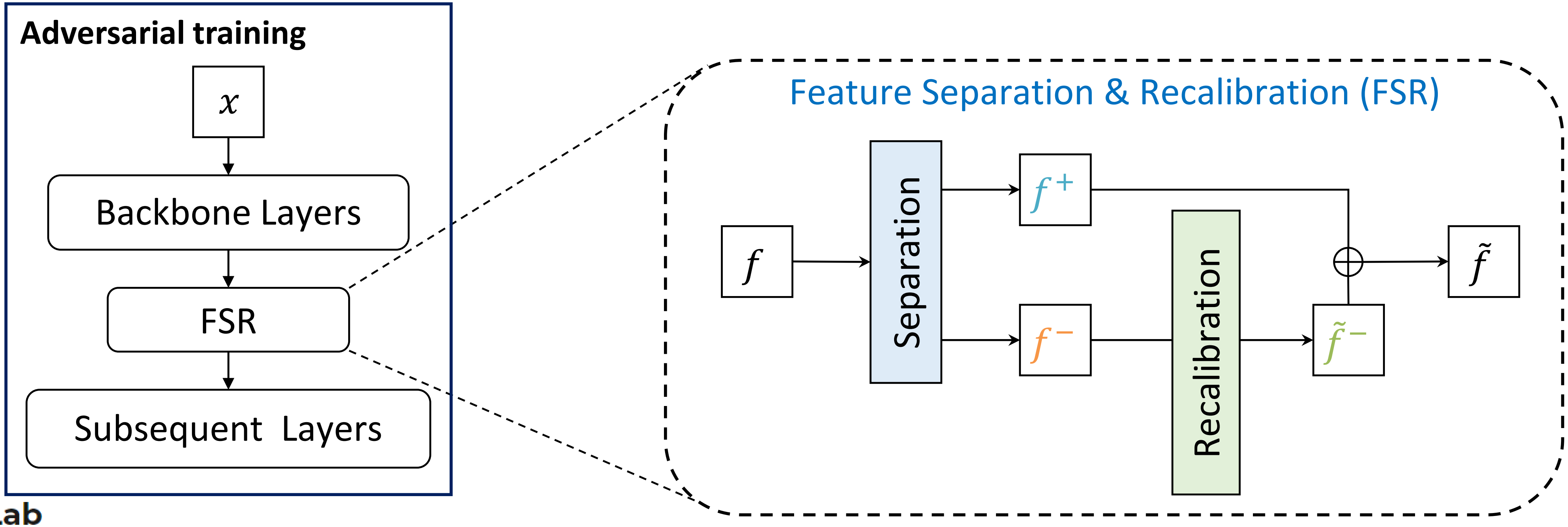
Feature Recalibration

- Conventional methods simply suppress the non-robust feature f^-
- This approach can ***neglect potentially useful cues*** in the non-robust feature



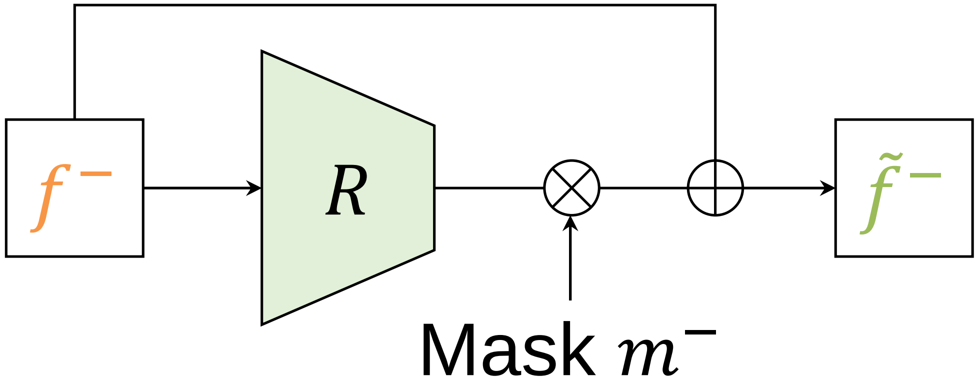
Feature Recalibration

- Recalibration: Recalibrates non-robust feature f^- to restore useful cues
- Recalibrated $\tilde{f}^-$: Activations with restored useful cues



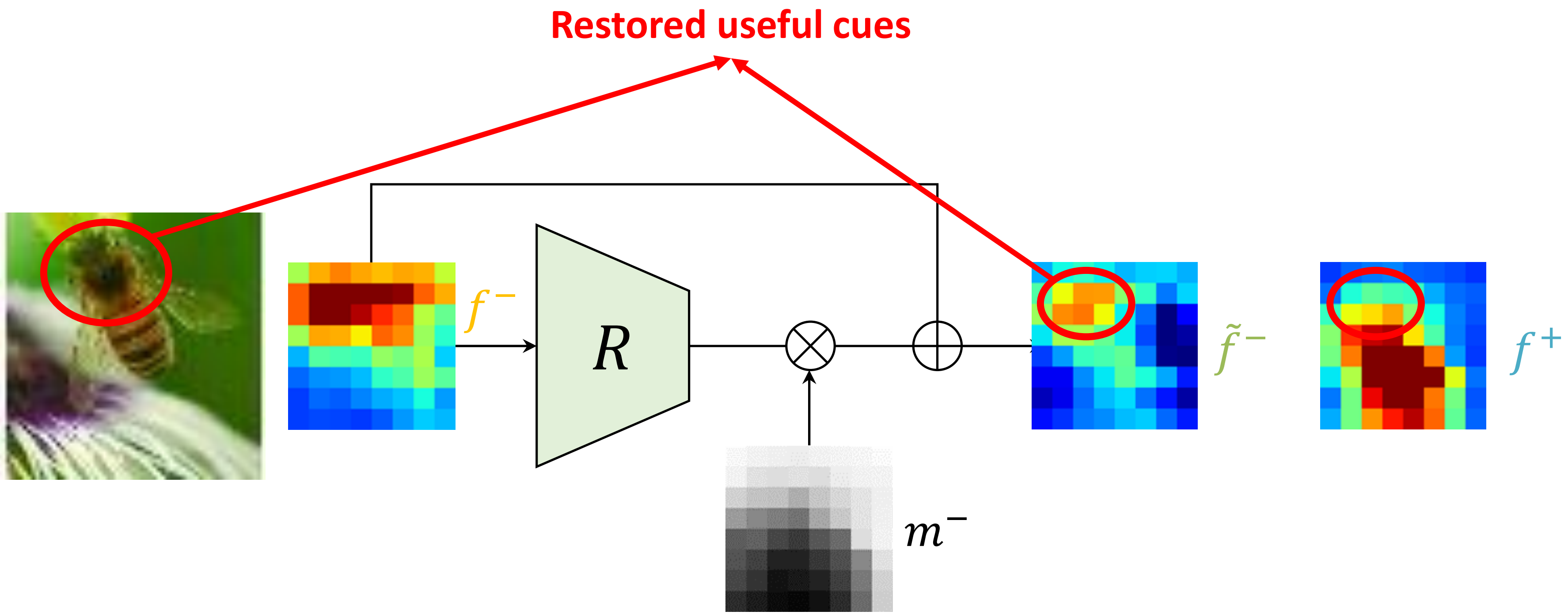
Feature Recalibration

- Recalibration Net R outputs recalibrating units
- We apply the recalibrating units on the non-robust feature f^-



Feature Recalibration

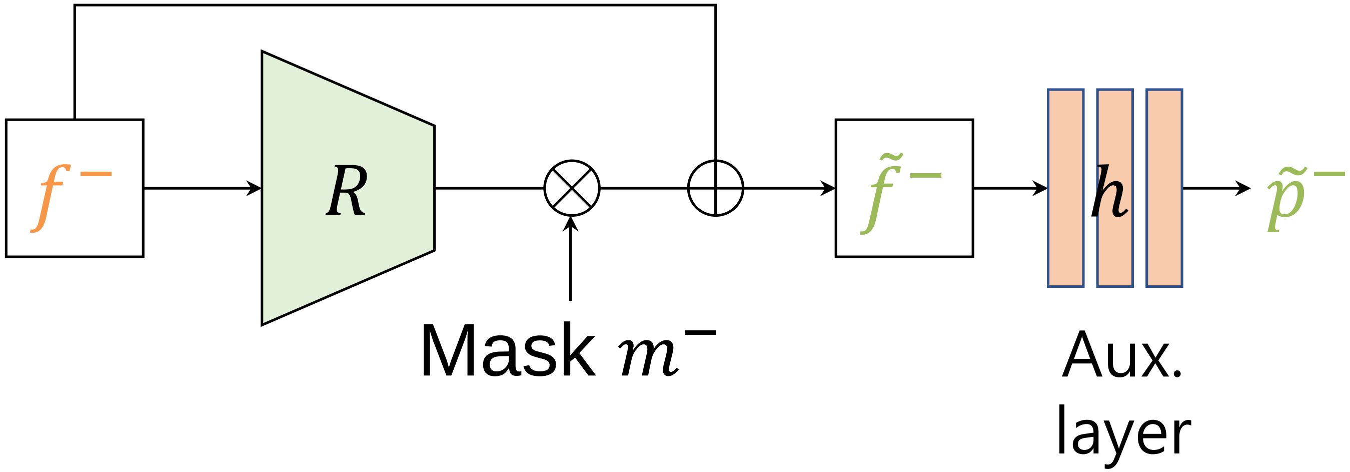
- Recalibration restores useful cues from non-robust feature
- These restored cues provide additional information for correct predictions



Feature Recalibration

- Guide the Rec. Net R to restore useful cues relevant to correct prediction

$$\mathcal{L}_{rec} = \underbrace{\mathcal{H}}_{\text{Cross-entropy loss}}(\underbrace{\tilde{p}^{-}}_{\text{Pred. logit}}, \underbrace{y}_{\text{GT label}})$$



Training

- Can be attached to any adversarial training (AT) technique with objective $\mathcal{L}_{cls}$

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{sep}\mathcal{L}_{sep} + \lambda_{rec}\mathcal{L}_{rec}$$

- Highly modularized
- Easy to plugin
- Trained in an end-to-end manner

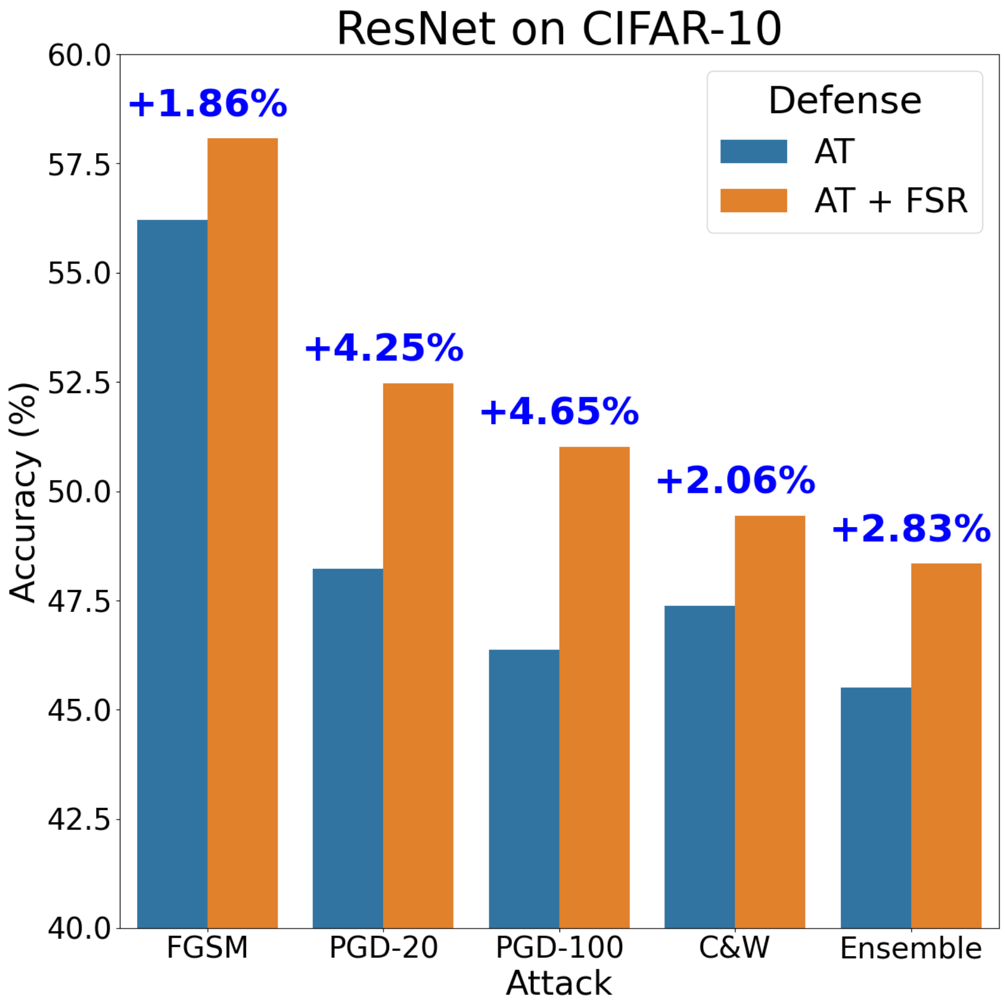
Experimental Evaluations

Experimental Setups

- Baselines
  - PGD adversarial training [1]
  - TRADES [2]
  - MART [3]
- Datasets
  - CIFAR-10/100
  - SVHN
  - Tiny ImageNet
- Models
  - ResNet18
  - VGG16
  - WideResNet-34-10

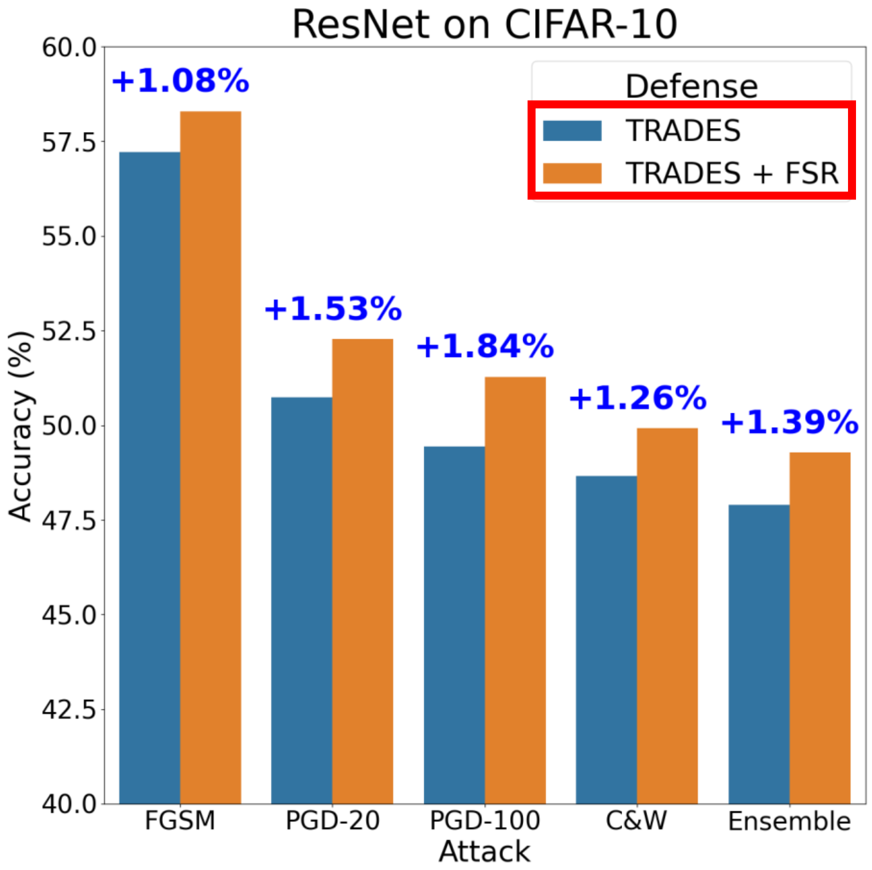
[1] Madry et al., Towards deep learning models resistant to adversarial attacks, ICLR 2018
[2] Zhang et al., Theoretically principled trade-off between robustness and accuracy, ICML 2019
[3] Wang et al., Improving adversarial robustness requires revisiting misclassified examples, ICLR 2019

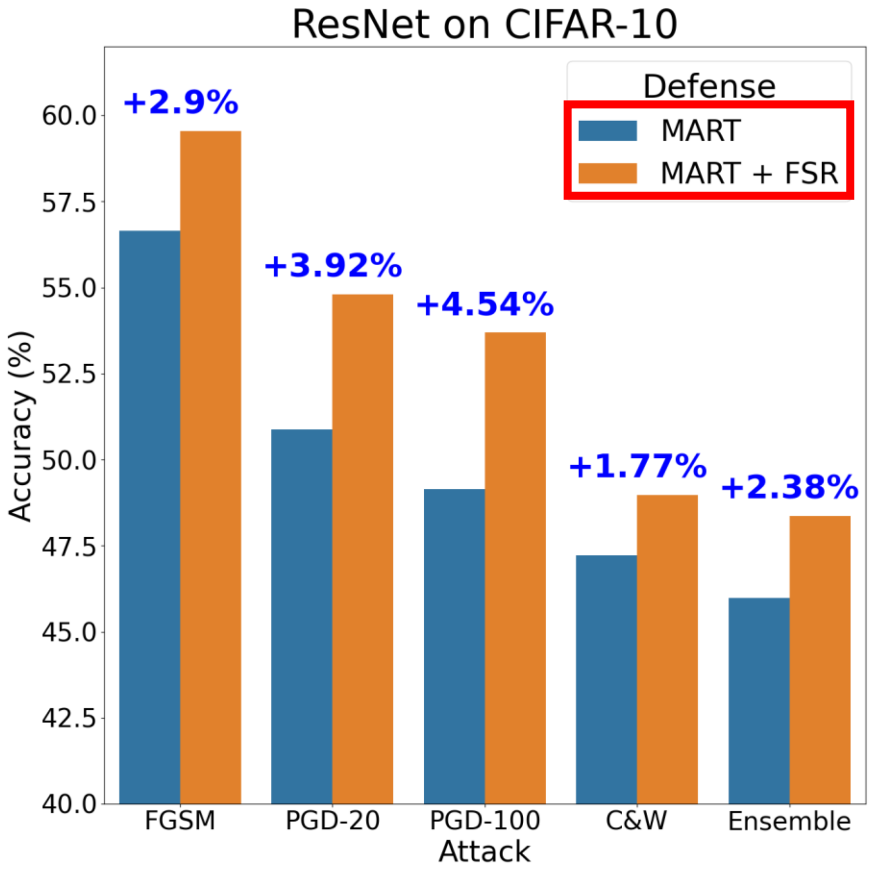
Improving Robustness of Adversarial Training



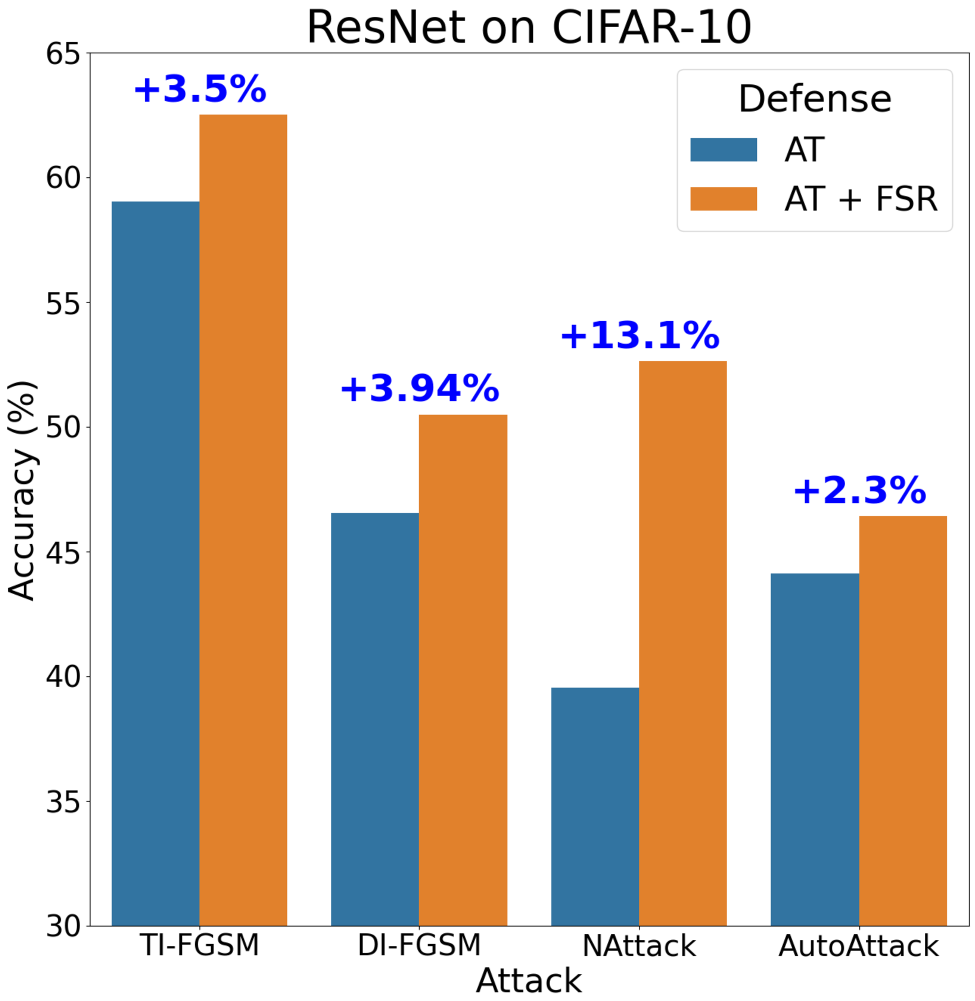
SGVR Lab

Improving Robustness of Adversarial Training | Different Baselines



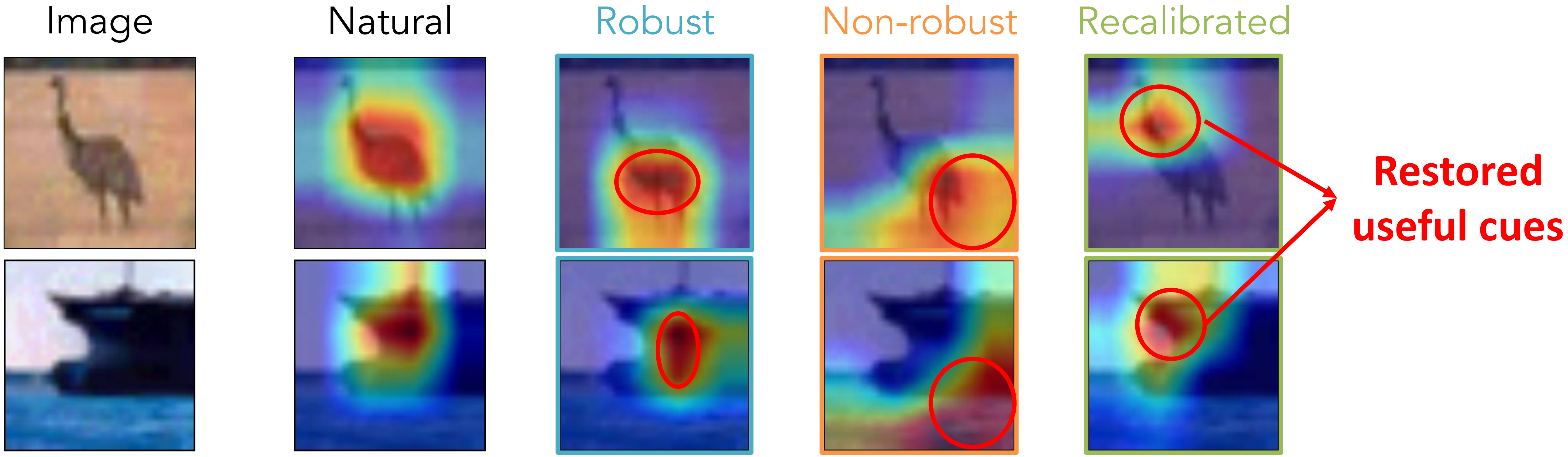


Robustness against Black-Box Attacks and AutoAttack



Robustness of Recalibrated Feature

Method	(a) Classification		(b) Weighted k-NN	
	Ensemble	AutoAttack	5-NN	20-NN
Robust f^+	47.89	45.82	66.21	61.58
Non-robust f^-	33.11	28.39	54.69	53.89
Recalibrated $\tilde{f}^-$	46.93	44.52	66.34	65.64
Combined $\tilde{f}(f^+ + \tilde{f}^-)$	48.34	46.41	70.91	65.88



Comparison w/ Conventional Methods

- Metric: Classification Accuracy (%)

Feature Deactivation or Suppression

Method	Ensemble	AutoAttack
AT [ICLR 2018]	45.51	44.11
FD [CVPR 2019]	45.82	44.57
CAS [ICLR 2021]	46.46	44.23
CIFS [ICML 2021]	47.26	43.94
FSR (Ours)	**48.34**	**46.41**

Take-home Messages

- FSR: Module to restore useful cues from disrupted features
- Highly modularized and easy-to-plugin
- Improves robustness of adversarial training-based techniques

Github Codes

github.com/wkim97/FSR

Project webpage

sgvr.kaist.ac.kr/~wjkim/FSR

Paper

https://arxiv.org/abs/2303.13846