JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

IMAGE & VIDEO SYSTEMS Lab. since 1998
Empowered by Deep Learning

Demystifying Causal Features on Adversarial Examples and Causal Inoculation for Robust Network by Adversarial Instrumental Variable Regression

Junho Kim*   Byung-Kwan Lee*   Yong Man Ro

{arkimjh, leebk, ymro} @ kaist.ac.kr
** Equal contribution*

Image and Video Systems Lab,

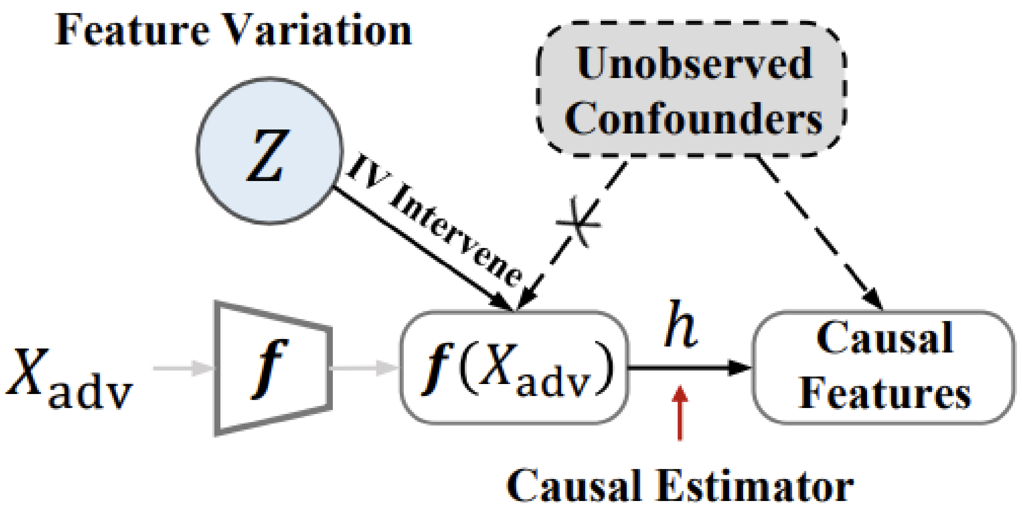
School of Electrical Engineering,

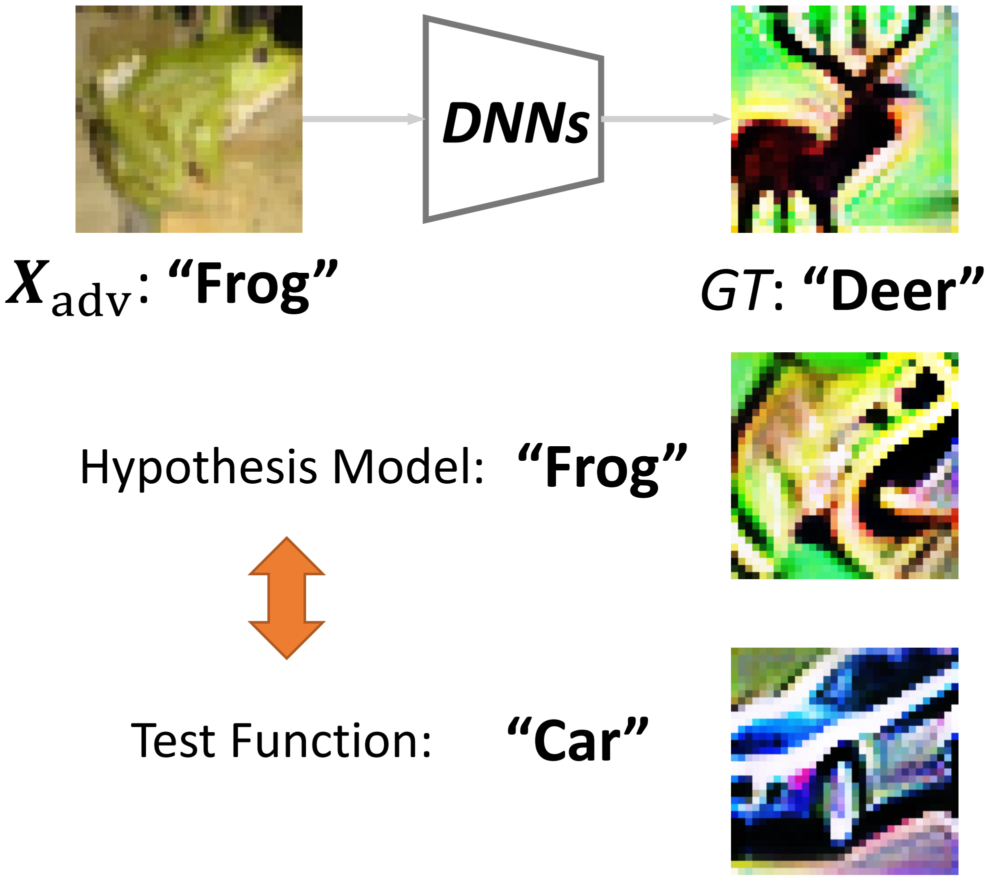
KAIST

https://github.com/ByungKwanLee/Causal-Adversarial-Instruments

https://paperswithcode.com/paper/demystifying-causal-features-on-adversarial

Quick Summary

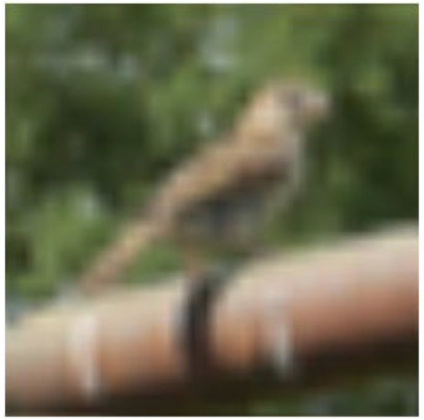




Necessity of Adversarial Robustness

- Adversarial examples, generated by carefully crafted perturbation, have attracted considerable attention in research fields.

x

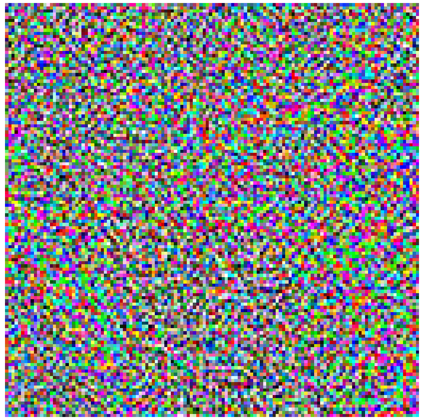


Clean example

"Bird"

$+$

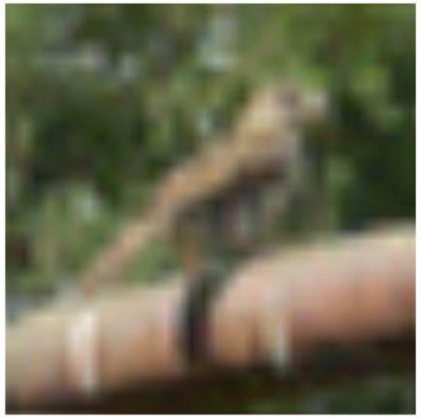
δ



Adversarial Perturbation

$=$

x_{adv}



Adversarial Example

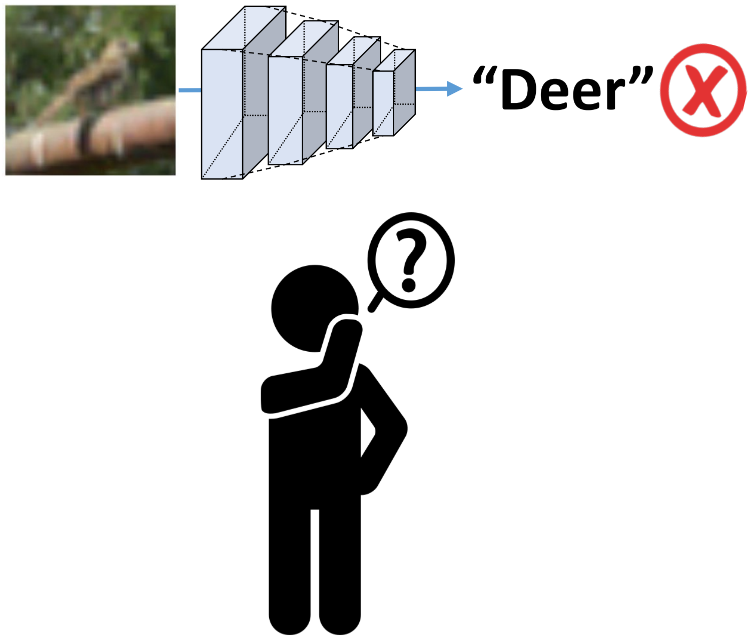
"Deer"

Potential Security Threats!

[Goodfellow et al. 2014]

Necessity of Adversarial Robustness

- But, it is still inexplicable the origin of adversarial examples, and it arouses arguments from various viewpoints, albeit comprehensive investigations.



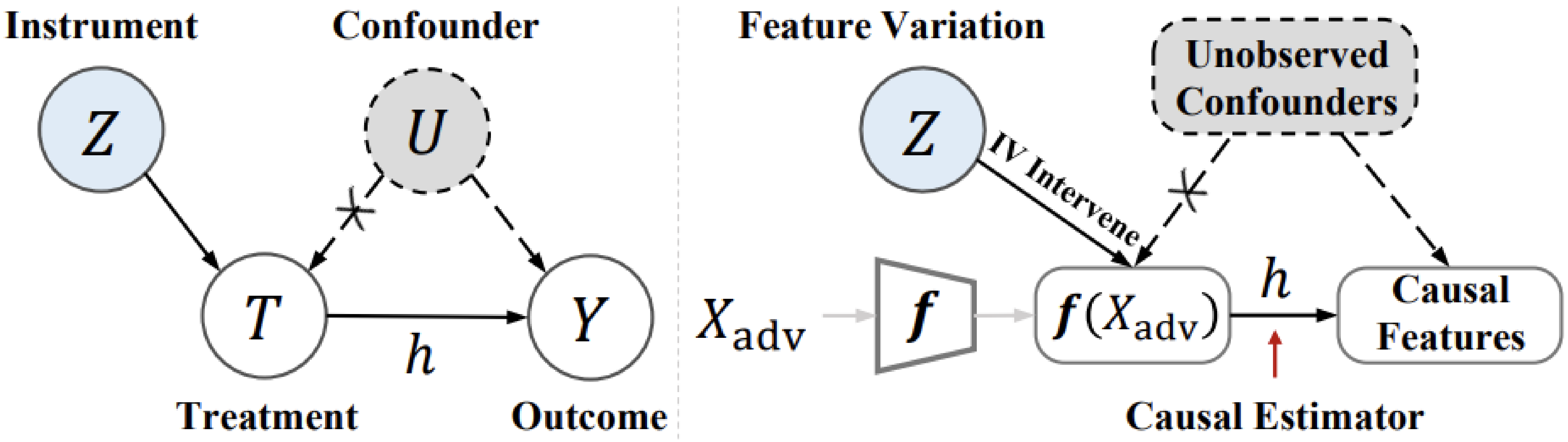
- Excessive linearity in a hyperplane?
- Aberration of statistical fluctuations?
- Induced from high-freq information?
- Existence of non-robust features?

[Goodfellow 2015; Shafahi et al. 2018; Yin et al. 2019; Kim et al. 2021]

Question

What is Inherent Causal Feature in Adversarial Examples?

Question



Backdoor Path: $T \leftarrow U \rightarrow Y$

Causal Path: $Z \rightarrow T \rightarrow Y$

Preliminary

Conditional Moment Restriction (CMR)

$$\mathbb{E}_{\underbrace{T}_{\text{Treatment}}}[\psi_T(h) \mid \underbrace{Z}_{\text{Instrument}}] = \mathbf{0},$$

where *a generalized residual function* $\psi_T(h)$ is denoted by

$$\psi_T(h) = \underbrace{Y}_{\text{Outcome}} - \underbrace{h}_{\text{Hypothesis Model}}(T)$$

Here, CMR serves as a key of finding **hypothesis model**.

Proposed Method

Adversarial Moment Restriction (AMR)

$$\mathbb{E}_{\underline{T}}[\psi_T(h) \mid \underline{Z}] = \mathbf{0}, \quad (Z = F_{\text{adv}} - F_{\text{natural}})$$

Adversarial Features Feature Variations

where *a generalized residual function* $\psi_T(h)$ is denoted by

$$\psi_T(h) = \underline{Y} - \underline{h}(T)$$

Causal Features Neural Networks

Also, AMR serves as a key of finding **neural networks** outputting **causal features**.

Proposed Method

Adversarial Moment Restriction (AMR)

$$\mathbb{E}_{\underbrace{T}_{\text{Adversarial Features}}}[\psi_T(h) \mid \underbrace{Z}_{\text{Feature Variations}}] = \mathbf{0}, \qquad \big(Z = F_{\text{adv}} - F_{\text{natural}} \big)$$

where *a generalized residual function* $\psi_T(h)$ is denoted by

$$\psi_T(h) = \underbrace{Y}_{\text{Causal Features}} - \underbrace{h}_{\text{Neural Networks}}(T)$$

*However, conventional "CMR" did not work well in finding **a non-parametric hypothesis model** with high-dimensional treatment due to ill-posed estimates [7, 20, 45, 78]*

Proposed Method

Generalized Method of Moments (GMM) can help AMR!

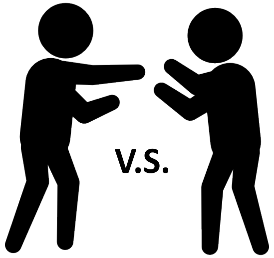
$$\mathbb{E}_{Z,T}[\psi_T(\underline{h}) \cdot \underline{g}(Z)] = 0 \qquad \left(\psi_T(h) = Y - h(T) \right)$$

Hypothesis Model	Test Function
(Causal) Satisfying CMR despite given counterfactual treatment	Generating counterfactual treatment against hypothesis model (Causal)
(Adversarial) Generating causal features given too much noisy feature	Interrupting generating causal features (Adversarial)

Implementation: Min-Max Optimization

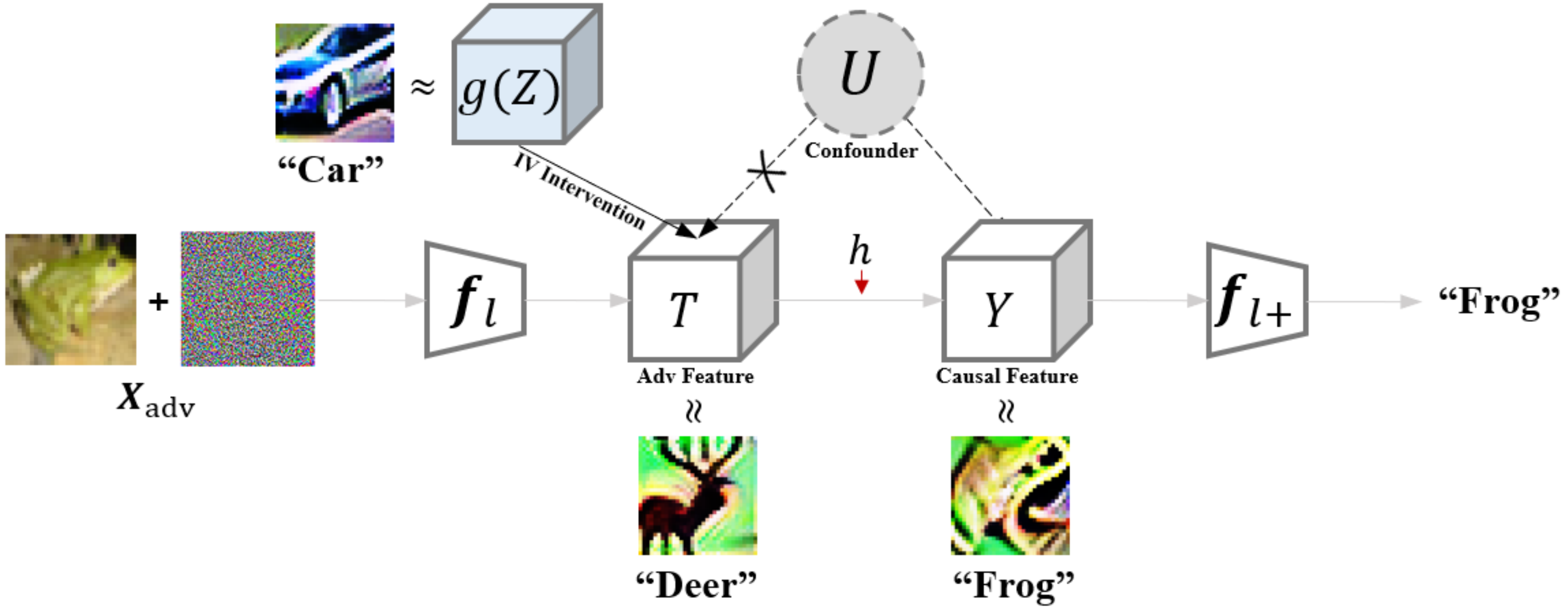
$$\min_{h \in \mathcal{H}} \max_{g \in \mathcal{G}} \mathbb{E}_{Z,T}[\psi_T(h) \cdot g(Z)]$$ (assuming the moment is semi-positive)

Causal Feature Generator   Counterfactual Feature Generator



[Dikkala et al. 2014]

Proposed Method



$$\min_{h} \max_{g} \mathbb{E}[\{\underset{?}{Y} - h(\underset{g(Z)}{\text{“Car”}})\} \times \underset{g(Z)}{\text{“Car”}}] \Rightarrow \min_{h} \max_{g} \mathbb{E}[\{\text{"frog"} - f_{l+}(\underset{h(g(Z))}{\text{“Frog”}})\} \times f_{l+}(\underset{g(Z)}{\text{“Car”}})]$$

None of Labels for Causal Features

Then, how?

Analysis on Causal Features

Feature Combinations

(i) Adversarial Feature (**Adv**)

$$F_{\text{adv}} = F_{\text{natural}} + Z$$

(ii) Counterfactual Feature (**CF**)

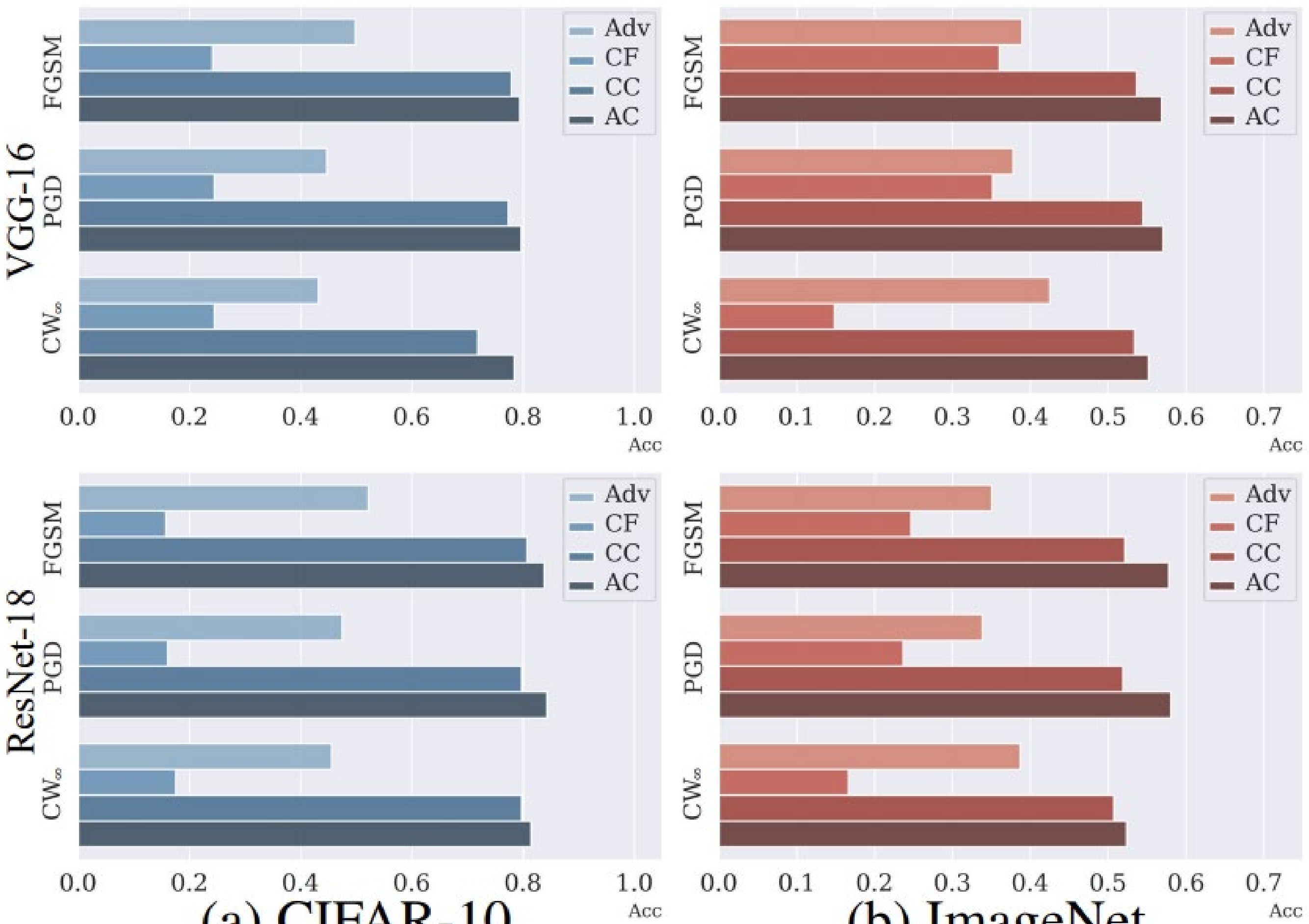
$$F_{\text{CF}} = F_{\text{natural}} + g(Z)$$

(iii) Counterfactual Causal Feature (**CC**)

$$F_{\text{CC}} = F_{\text{natural}} + (h \circ g)(Z)$$

(iv) Adversarial Causal Feature (**AC**)

$$F_{\text{AC}} = F_{\text{natural}} + h(Z)$$



(a) CIFAR-10 (b) ImageNet

Analysis on Causal Features

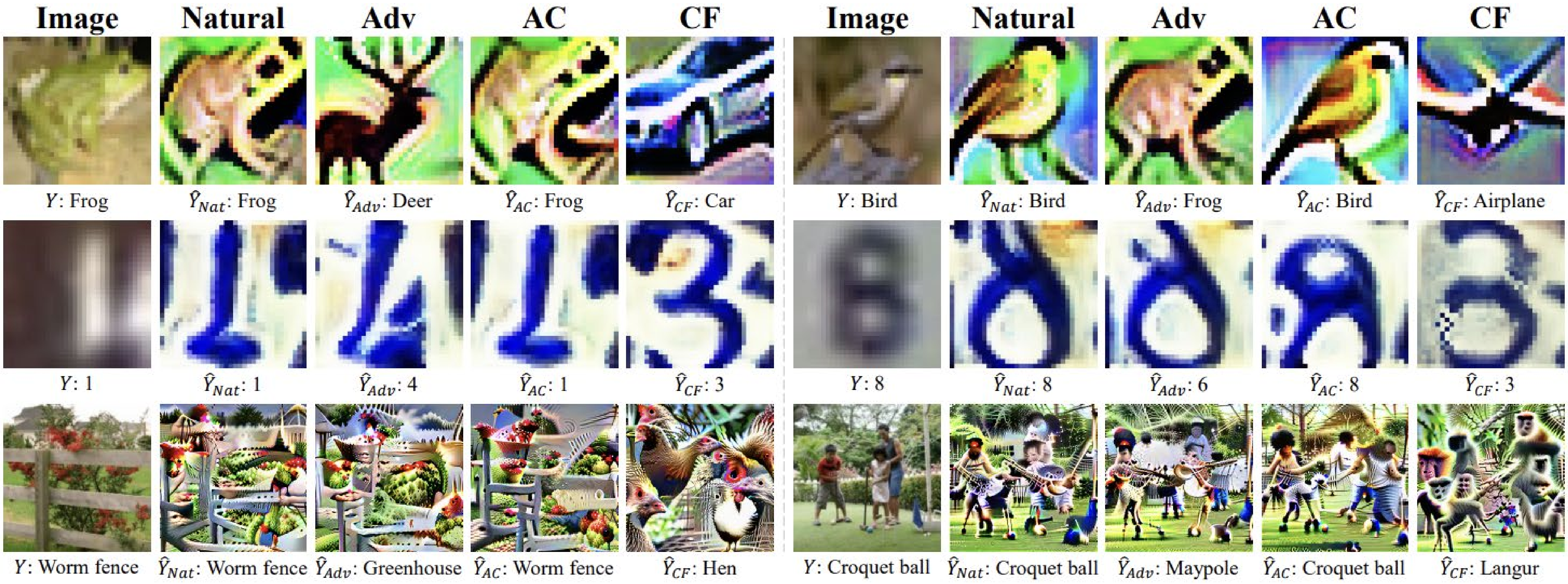


Fig 3. Comparison results of feature visualization for the defined feature variations.

[Olah et al. 2017]

Inoculating Causal FEatures

We can now implant CAusal FEatures to boost adversarial robustness!

$$\min_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{S}} \left[\max_{\|\epsilon\|_\infty \in \gamma} \mathcal{L}_{\mathrm{Defense}} + \mathcal{D}_{\mathrm{KL}}\left(f_{l+}\left(\hat{F}_{\mathrm{AC}}\right) | f_{l+}(F_{\mathrm{adv}})\right) \right]$$

Validating CAFE

KAIST IMAGE & VIDEO SYSTEMS Lab. since 1998 Empowered by Deep Learning

Defense Baselines

- AT
- TRADES
- MART
- HELP
- AWP

Attack Baselines

- FGSM
- PGD
- CW
- AP / DLR
- AA

	Method	CIFAR-10							SVHN							Tiny-ImageNet						
		Natural	FGSM	PGD	CW_∞	AP	DLR	AA	Natural	FGSM	PGD	CW_∞	AP	DLR	AA	Natural	FGSM	PGD	CW_∞	AP	DLR	AA
VGG	ADV	**78.5**	49.8	44.8	42.6	43.2	42.9	40.7	**91.9**	64.8	52.1	48.9	48.0	48.5	45.2	**53.2**	25.3	21.5	21.0	20.2	20.8	19.6
	ADV_{CAFE}	78.4	**52.2**	**47.9**	**44.1**	**46.4**	**44.5**	**42.7**	91.5	**67.0**	**55.3**	**50.0**	**51.3**	**49.6**	**46.1**	52.6	**26.0**	**22.8**	**22.1**	**21.8**	**22.0**	**21.0**
	TRADES	**79.5**	50.4	45.7	43.2	44.4	42.9	41.8	**91.9**	66.4	53.6	49.1	49.1	47.7	45.2	**52.8**	25.9	22.5	21.9	21.5	21.8	20.7
	$\text{TRADES}_{\text{CAFE}}$	77.0	**51.6**	**47.9**	**44.0**	**47.0**	**43.9**	**42.7**	90.3	**67.8**	**56.1**	**50.0**	**53.6**	**49.1**	**47.5**	52.1	**26.5**	**23.6**	**22.6**	**22.5**	**22.6**	**21.6**
	MART	**79.7**	52.4	47.2	43.4	45.5	43.8	42.0	**92.6**	66.6	54.2	47.9	49.6	47.1	44.4	**53.1**	25.0	21.5	21.2	20.4	21.0	19.9
	$\text{MART}_{\text{CAFE}}$	78.3	**54.2**	**49.7**	**43.9**	**48.1**	**44.5**	**42.7**	91.3	**67.6**	**57.3**	**49.5**	**54.2**	**48.3**	**46.4**	53.0	**25.6**	**22.3**	**21.6**	**21.3**	**21.5**	**20.5**
	AWP	**78.0**	51.7	48.2	43.5	47.2	43.4	42.6	90.8	65.5	56.6	50.4	54.0	49.7	48.6	52.6	28.0	25.7	23.6	24.8	23.5	22.8
	AWP_{CAFE}	77.4	**54.8**	**51.4**	**44.2**	**50.2**	**44.9**	**43.5**	**91.9**	**67.9**	**58.6**	**51.2**	**55.9**	**51.1**	**49.7**	**52.9**	**28.8**	**26.4**	**24.2**	**25.6**	**24.1**	**23.4**
	HELP	**77.4**	51.8	48.3	43.9	47.3	43.9	42.9	91.2	65.8	56.6	50.9	53.9	50.2	48.8	**53.0**	28.3	25.9	23.9	25.1	23.8	23.1
	$\text{HELP}_{\text{CAFE}}$	75.6	**54.4**	**51.4**	**44.6**	**50.4**	**44.8**	**43.7**	**91.5**	**67.3**	**58.5**	**51.6**	**56.2**	**51.4**	**50.0**	52.6	**29.4**	**27.1**	**24.7**	**26.4**	**24.4**	**23.9**
ResNet	ADV	82.0	52.1	46.5	44.8	44.8	44.8	43.0	**92.8**	70.4	55.4	51.3	50.9	51.0	47.5	**57.2**	27.3	24.2	23.2	22.8	23.2	21.8
	ADV_{CAFE}	**82.6**	**55.9**	**50.7**	**47.6**	**49.0**	**47.7**	**46.2**	92.5	**73.6**	**58.9**	**53.8**	**54.9**	**52.6**	**49.8**	56.3	**28.6**	**25.7**	**24.7**	**24.4**	**24.6**	**23.5**
	TRADES	**83.0**	55.0	49.8	47.5	48.3	47.3	46.1	**93.2**	72.8	57.7	52.6	53.0	51.5	48.9	**56.5**	28.4	25.3	24.4	24.2	24.3	23.2
	$\text{TRADES}_{\text{CAFE}}$	80.7	**56.6**	**51.4**	**48.5**	**50.4**	**48.3**	**46.7**	91.3	**73.9**	**59.6**	**54.1**	**56.7**	**53.2**	**51.3**	54.5	**29.6**	**27.4**	**26.3**	**26.5**	**26.2**	**25.4**
	MART	**83.5**	56.1	50.1	47.1	48.3	47.0	45.5	**93.7**	74.2	58.3	51.7	53.2	50.8	47.8	**57.1**	27.4	24.2	23.2	22.9	23.2	22.2
	$\text{MART}_{\text{CAFE}}$	82.1	**57.3**	**51.9**	**48.1**	**50.2**	**48.0**	**46.2**	92.2	**74.9**	**61.0**	**53.4**	**57.3**	**51.8**	**49.7**	55.9	**28.6**	**25.9**	**24.6**	**24.7**	**24.5**	**23.5**
	AWP	81.2	55.3	51.6	48.0	50.5	47.8	46.9	92.2	71.1	59.8	54.3	56.8	53.6	52.0	56.2	30.5	28.5	26.2	27.6	26.2	25.5
	AWP_{CAFE}	**81.5**	**57.8**	**54.2**	**49.4**	**52.9**	**49.0**	**47.8**	**93.4**	**74.0**	**60.9**	**55.0**	**57.8**	**54.8**	**52.7**	**56.6**	**31.4**	**29.2**	**27.1**	**28.4**	**27.0**	**26.5**
	HELP	80.5	55.8	52.1	48.4	51.1	48.5	47.4	92.6	72.0	59.8	54.4	56.6	53.9	52.0	**56.1**	31.0	28.6	26.3	27.7	26.3	25.7
	$\text{HELP}_{\text{CAFE}}$	**80.6**	**57.8**	**54.5**	**49.4**	**53.1**	**49.5**	**48.5**	**92.9**	**73.9**	**61.3**	**55.3**	**58.8**	**54.6**	**52.8**	55.4	**32.0**	**29.7**	**27.4**	**29.2**	**27.8**	**27.3**
WRN	ADV	84.3	54.5	48.7	47.8	47.0	47.9	45.6	**94.0**	71.8	56.7	53.2	51.9	52.8	49.0	**60.9**	29.8	25.5	25.8	24.2	26.0	23.9
	ADV_{CAFE}	**85.7**	**58.5**	**53.3**	**51.3**	**51.8**	**51.5**	**49.5**	93.7	**75.7**	**59.1**	**54.9**	**54.0**	**54.1**	**50.2**	60.6	**31.1**	**27.3**	**27.2**	**25.8**	**27.4**	**25.4**
	TRADES	**86.3**	57.1	52.1	50.8	50.6	50.7	49.0	**93.8**	74.0	58.1	53.9	53.0	53.4	49.9	**60.8**	30.5	26.4	26.7	25.0	26.8	24.6
	$\text{TRADES}_{\text{CAFE}}$	83.7	**58.6**	**54.5**	**52.0**	**53.2**	**52.0**	**50.1**	92.4	**75.6**	**61.0**	**55.7**	**58.0**	**58.0**	**53.0**	60.3	**31.7**	**28.2**	**28.3**	**27.0**	**28.5**	**26.5**
	MART	**86.5**	58.5	52.6	50.0	50.7	49.9	48.0	**94.2**	75.0	58.0	53.1	52.8	52.8	48.9	**60.7**	29.9	25.6	25.9	24.0	25.5	23.6
	$\text{MART}_{\text{CAFE}}$	85.7	**59.8**	**54.6**	**51.4**	**52.7**	**50.9**	**49.3**	93.0	**76.5**	**61.9**	**54.9**	**57.2**	**53.8**	**50.7**	60.4	**31.2**	**27.5**	**26.8**	**25.5**	**27.0**	**25.1**
	AWP	83.7	58.0	54.7	51.3	53.7	51.2	50.1	93.2	73.4	60.8	55.9	57.5	55.5	53.6	**61.9**	35.5	32.8	31.0	31.6	31.1	29.6
	AWP_{CAFE}	**84.6**	**60.6**	**56.9**	**52.4**	**55.5**	**52.3**	**51.1**	**94.2**	**76.9**	**62.7**	**57.5**	**59.2**	**57.1**	**54.6**	61.4	**36.6**	**34.2**	**32.3**	**33.2**	**32.5**	**30.8**
	HELP	**83.8**	58.6	54.9	51.6	53.8	51.6	50.3	93.5	73.4	60.8	56.5	57.6	56.1	54.0	**61.8**	35.9	33.0	31.3	31.8	31.3	29.8
	$\text{HELP}_{\text{CAFE}}$	83.1	**60.5**	**57.1**	**52.7**	**56.0**	**52.6**	**51.3**	**94.0**	**76.6**	**62.6**	**57.7**	**58.8**	**57.2**	**55.0**	61.1	**37.0**	**34.7**	**32.6**	**33.8**	**32.8**	**31.2**

Table 2. Results of adversarial robustness using CAFE in existing defense methods

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

[Madry et al. 2017;Zhang et al. 2019; Wang et al. 2020; Wu et al. 2020; Rade et al. 2022;]

Thank you