Exploring the Connection between **Robust** and **Generative** Models

github.com/senad96/Robust-Generative

Senad Beadini
Computer Science Department
Sapienza, University of Rome*

Iacopo Masi
Computer Science Department
Sapienza, University of Rome

* Now at Eustema S.p.a.

Exploring the Connection between Robust and Generative Models

Why Robust Models behave as Generatives?

Generative Models

Generative Models

$$\{\mathbf{x}_i\}$$

Generative Models

$$\{\mathbf{x}_i\} \sim p_{\text{data}}(\mathbf{x})$$

Generative Models

$$\mathbf{x}' \sim p(\mathbf{x};\boldsymbol{\theta}) \quad \{\mathbf{x}_i\} \sim p_{\text{data}}(\mathbf{x})$$

Generative Models

$$\mathbf{x}' \sim p(\mathbf{x};\boldsymbol{\theta}) \quad \{\mathbf{x}_i\} \sim p_{\text{data}}(\mathbf{x})$$

GAN

Generative Models

$$\mathbf{x}' \sim p(\mathbf{x};\boldsymbol{\theta}) \quad \{\mathbf{x}_i\} \sim p_{\text{data}}(\mathbf{x})$$

GAN

VAE

Generative Models

$$\mathbf{x}' \sim p(\mathbf{x};\boldsymbol{\theta}) \quad \{\mathbf{x}_i\} \sim p_{\text{data}}(\mathbf{x})$$

GAN

VAE

Autoregressive, Normalizing Flows, Invertible NN, and EBM

Generative Models

$$\mathbf{x}' \sim p(\mathbf{x}; \boldsymbol{\theta}) \quad \{\mathbf{x}_i\} \sim p_{\text{data}}(\mathbf{x})$$

GAN

VAE

Autoregressive, Normalizing Flows, Invertible NN, and EBM

Diffusion Models

Generative Models

$\mathbf{x}' \sim p(\mathbf{x}; \boldsymbol{\theta}) \quad \{\mathbf{x}_i\} \sim p_{\text{data}}(\mathbf{x})$

GAN

VAE

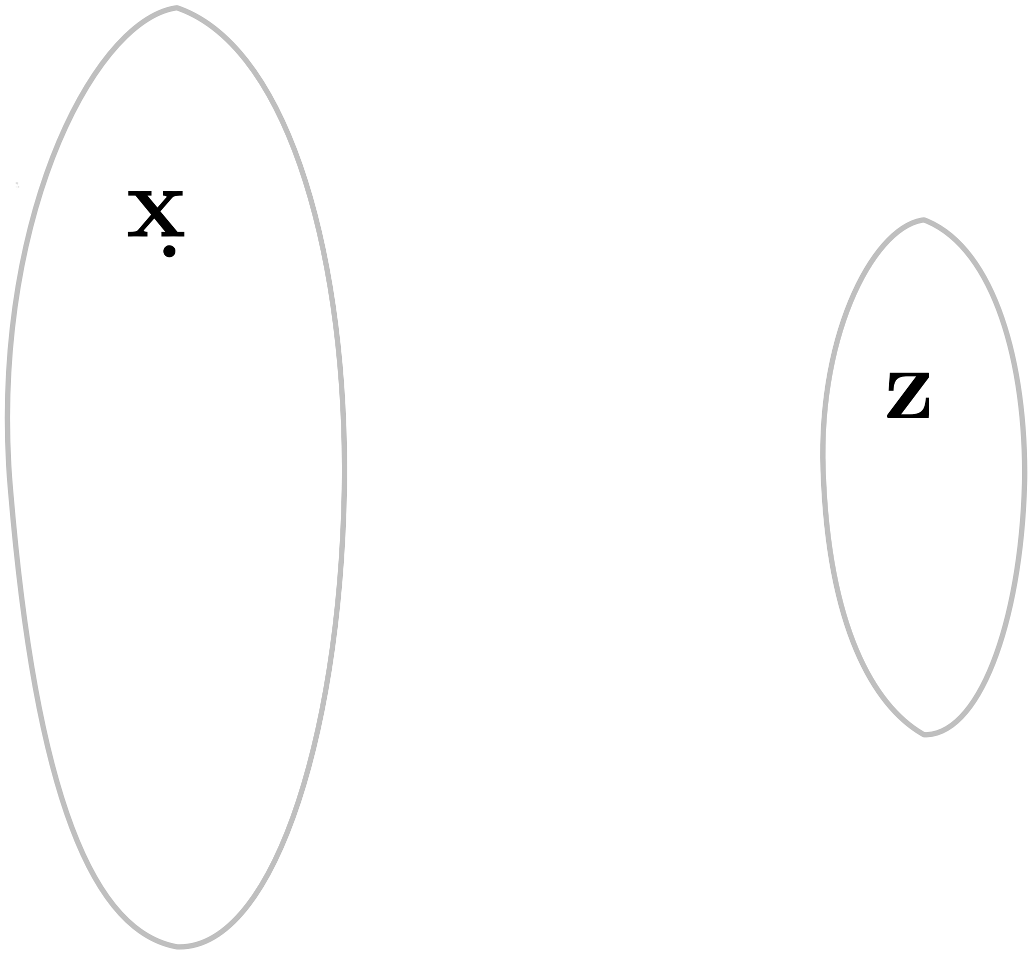
Autoregressive, Normalizing Flows, Invertible NN, and EBM

Diffusion Models

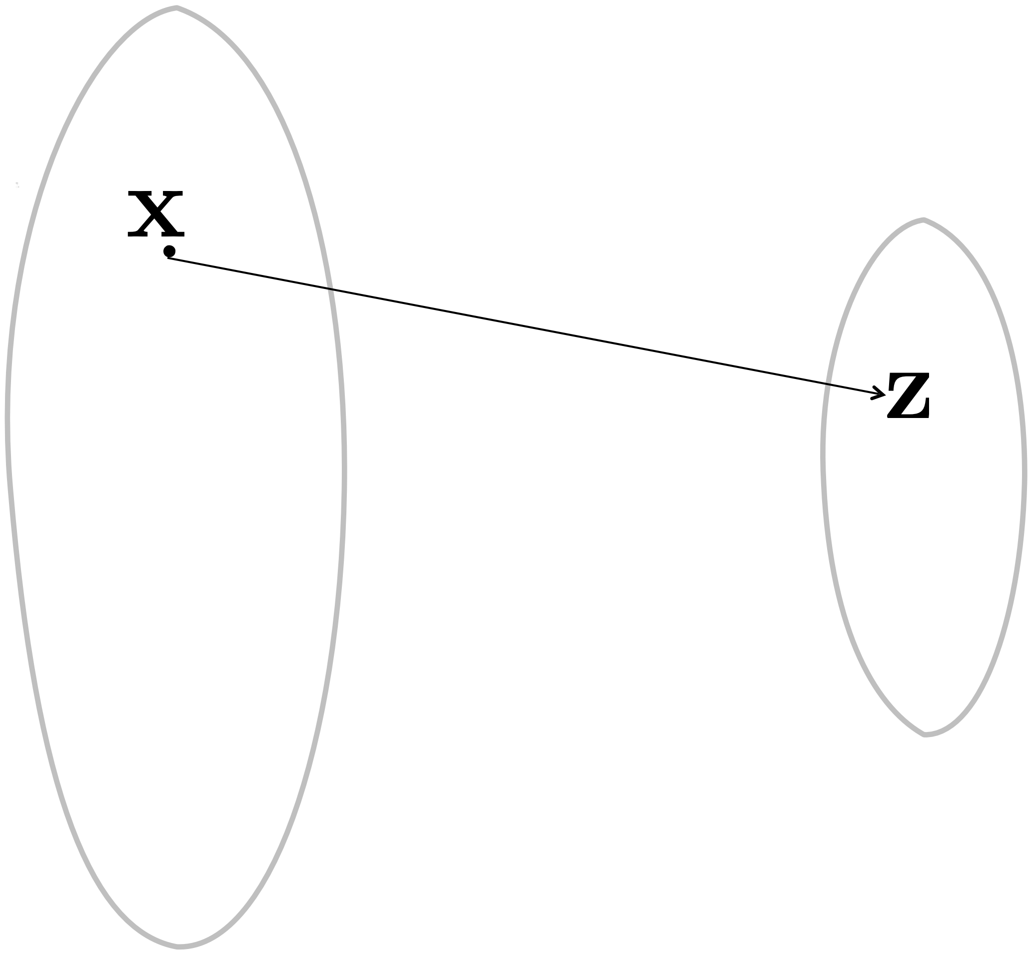
$p(z|\mathbf{x})$

Inverting a discriminative, robust model

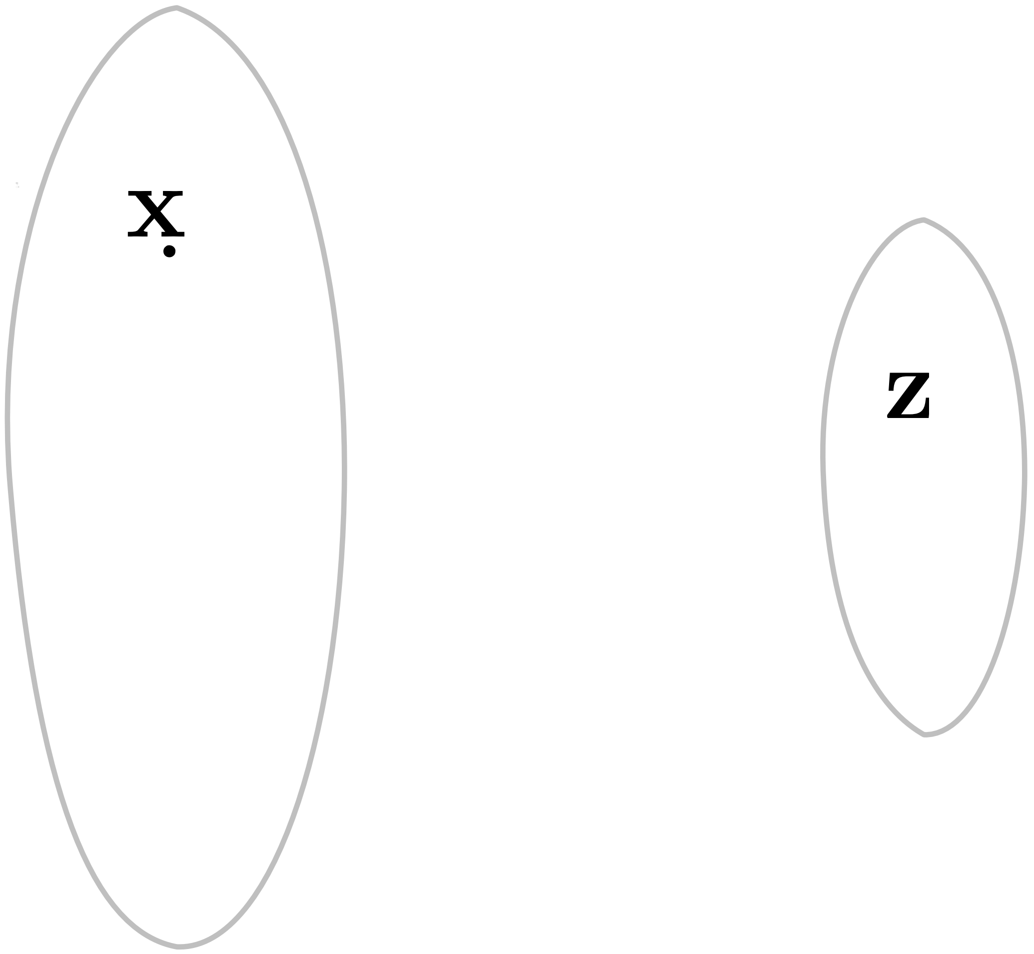
Robust Model $p(z|\mathbf{x})$



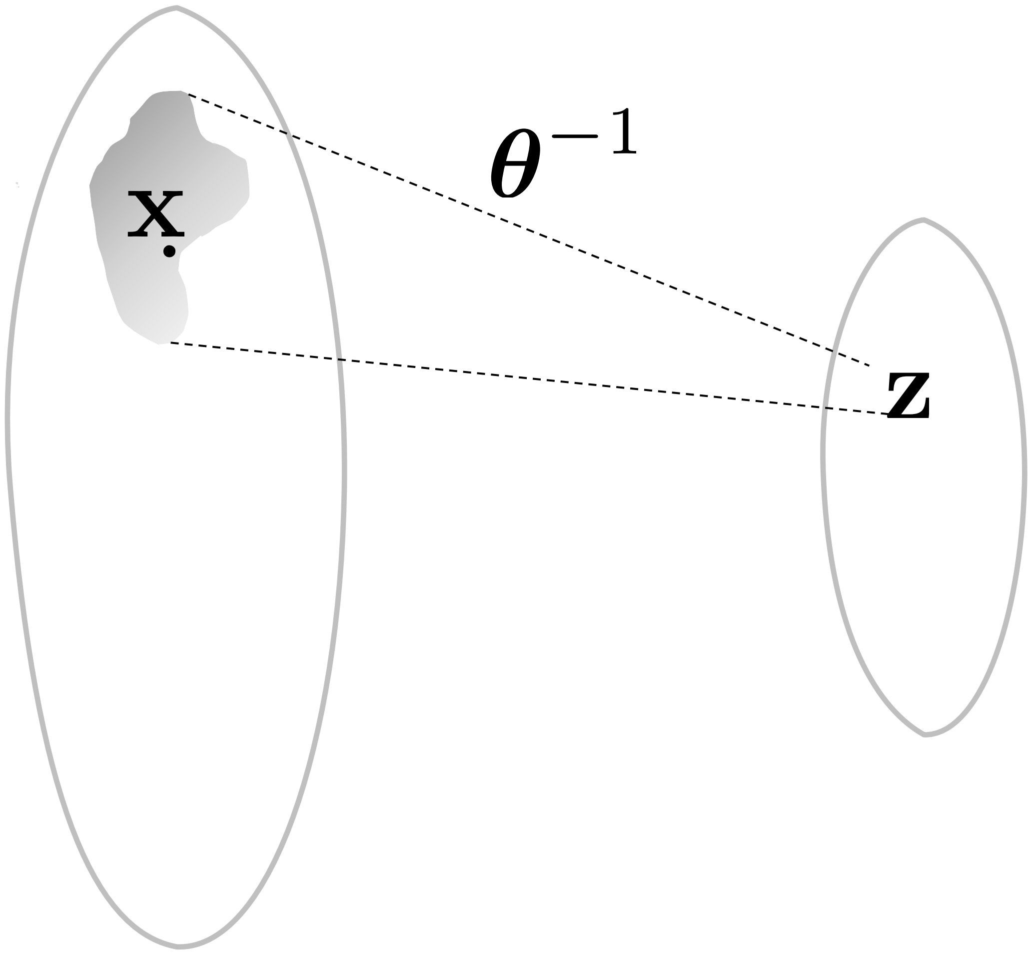
Robust Model $p(z|\mathbf{x})$



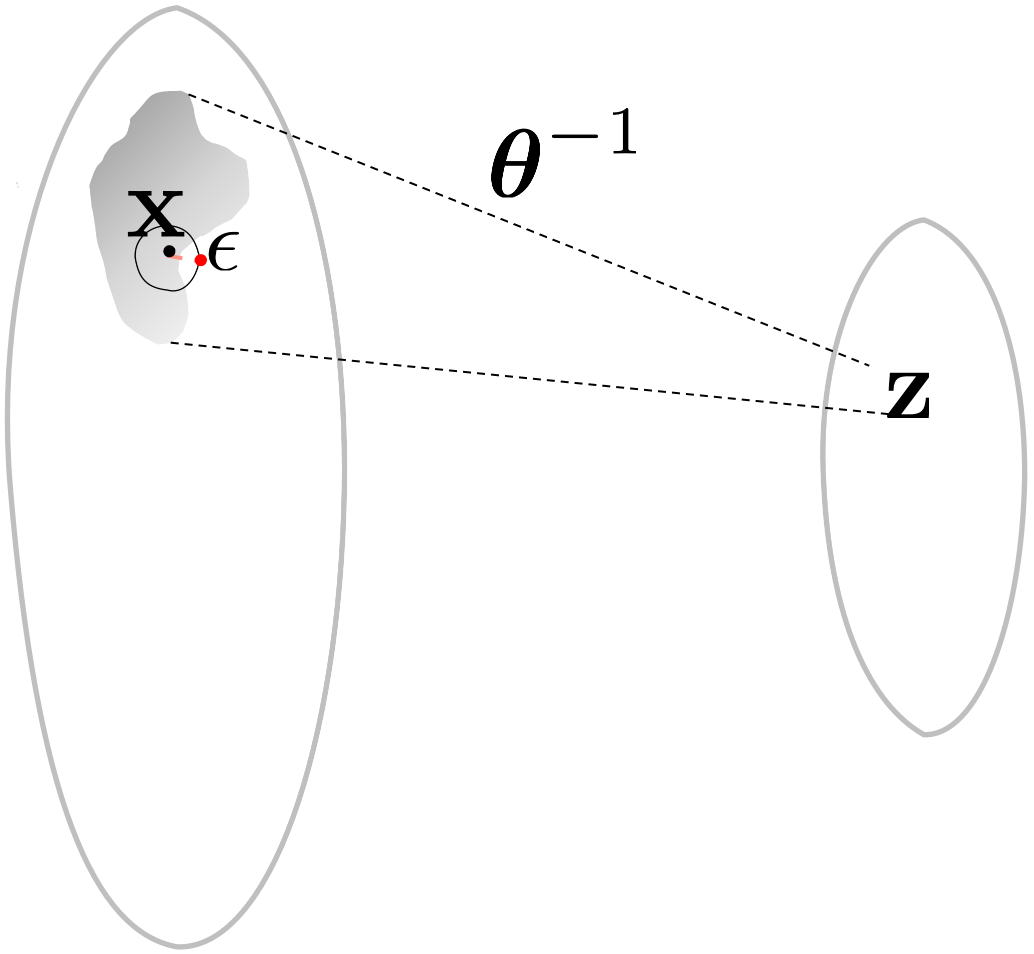
Robust Model $p(z|\mathbf{x})$



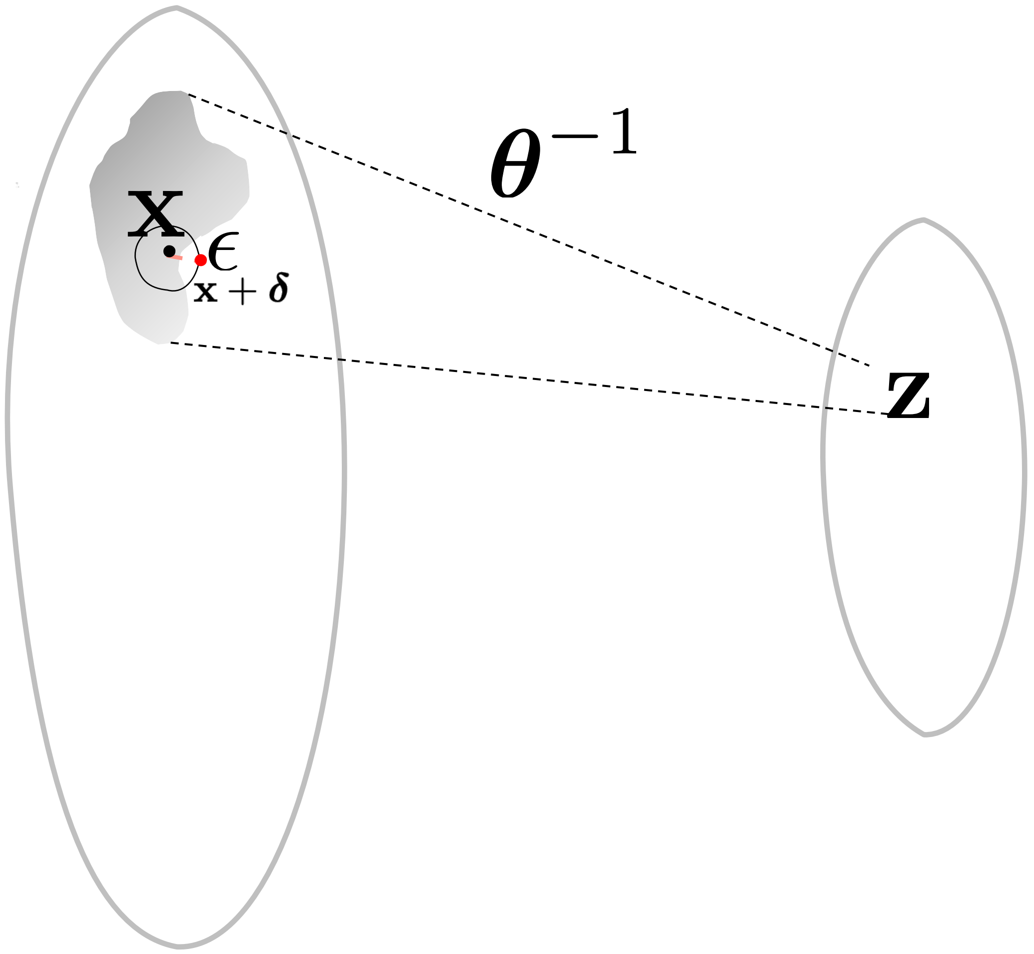
Robust Model $p(z|\mathbf{x})$



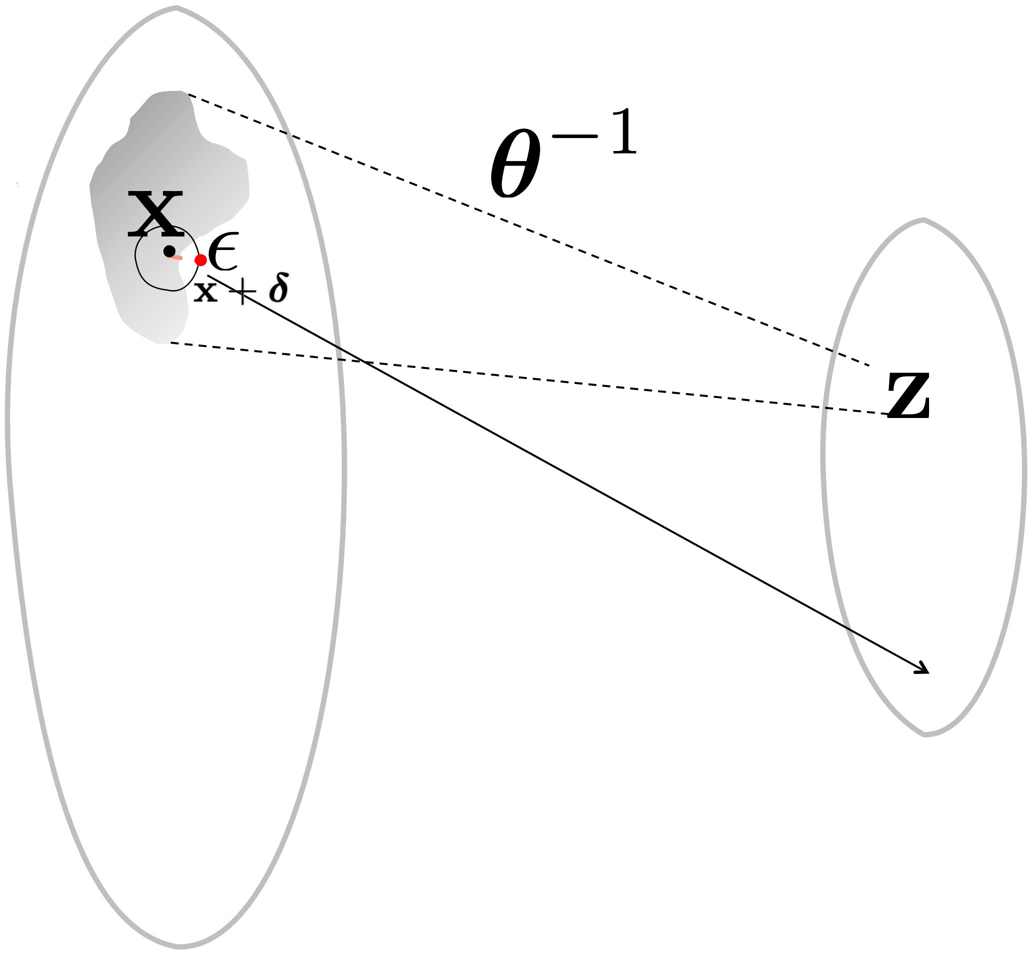
Robust Model $p(z|\mathbf{x})$



Robust Model $p(z|\mathbf{x})$

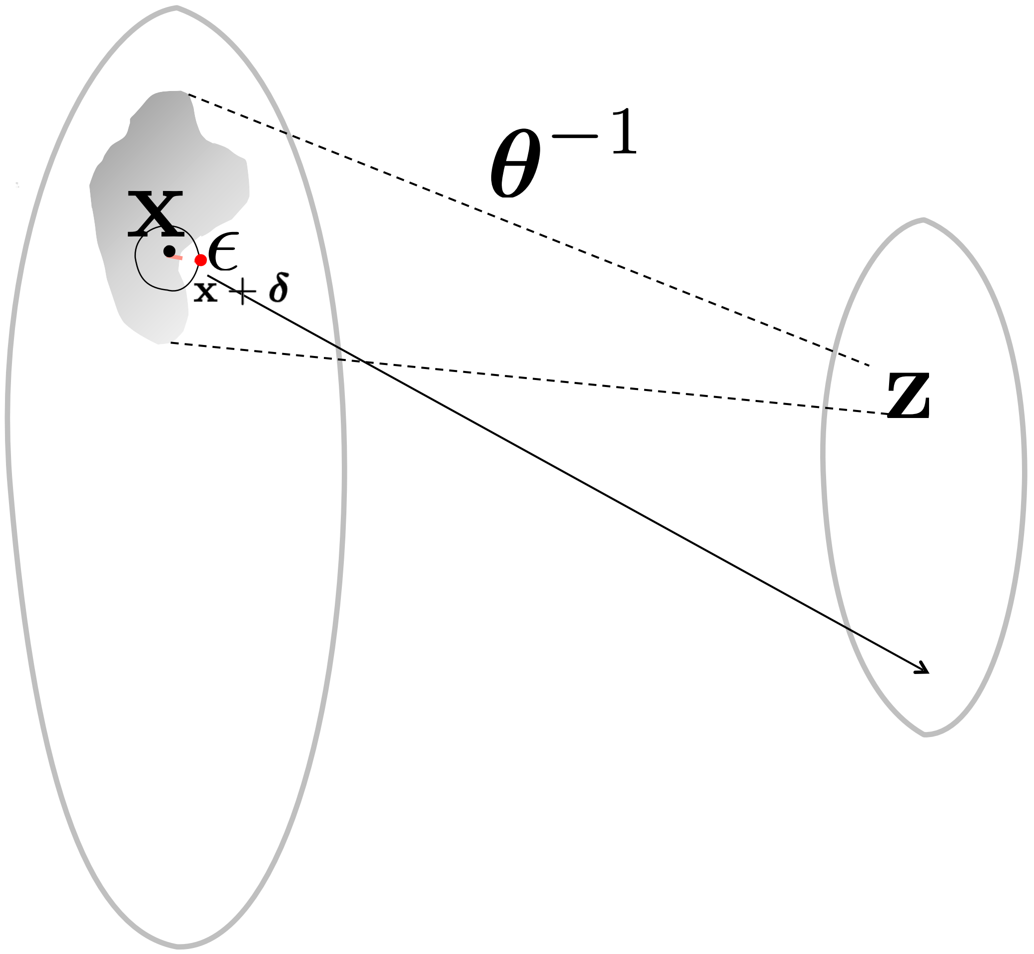


Robust Model $p(z|\mathbf{x})$



$$\boldsymbol{\delta}^{\star} = \arg\max_{||\boldsymbol{\delta}||_p < \epsilon} \ell\left(\boldsymbol{\theta}\left(\mathbf{x}+\boldsymbol{\delta}\right), y\right)$$

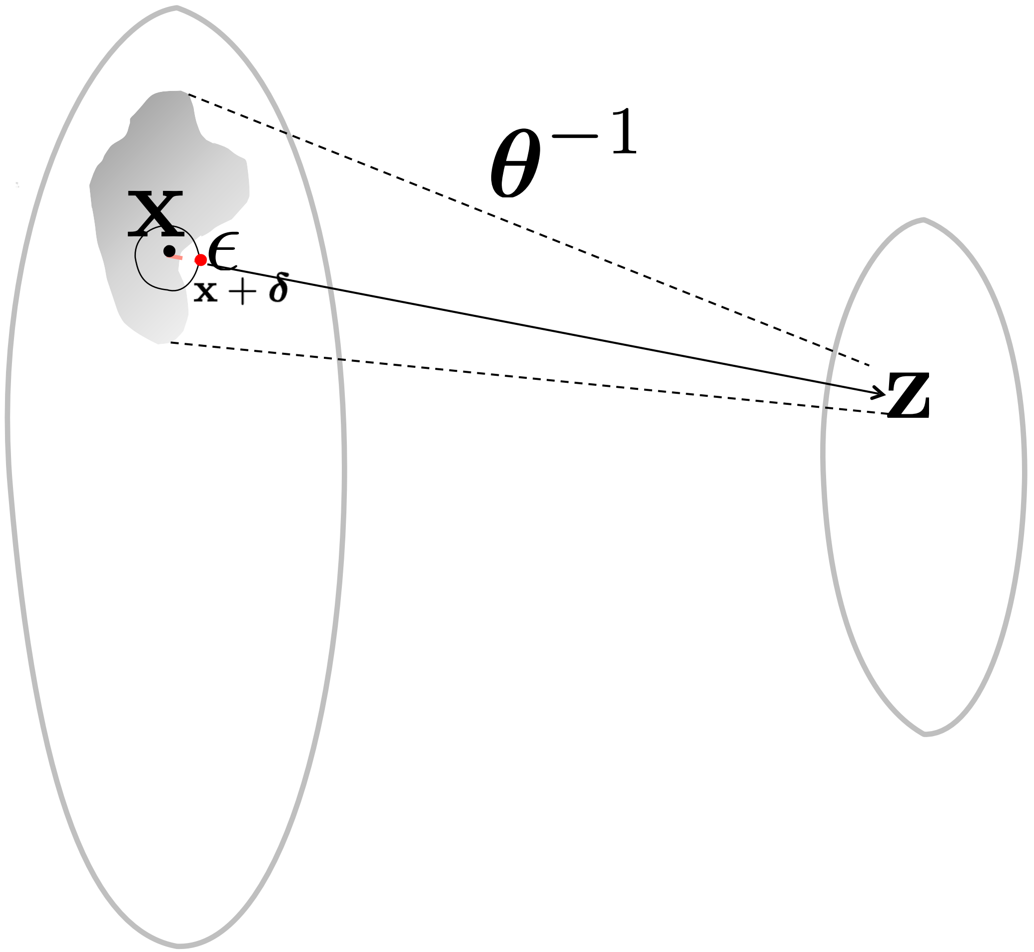
Robust Model $p(z|\mathbf{x})$



$$\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} \ell\big(\boldsymbol{\theta}(\mathbf{x}+\boldsymbol{\delta}^\star), y\big) \quad \text{where}$$

$$\boldsymbol{\delta}^\star = \arg\max_{||\boldsymbol{\delta}||_p<\epsilon} \ell\big(\boldsymbol{\theta}\left(\mathbf{x}+\boldsymbol{\delta}\right), y\big)$$

Robust Model $p(z|\mathbf{x})$

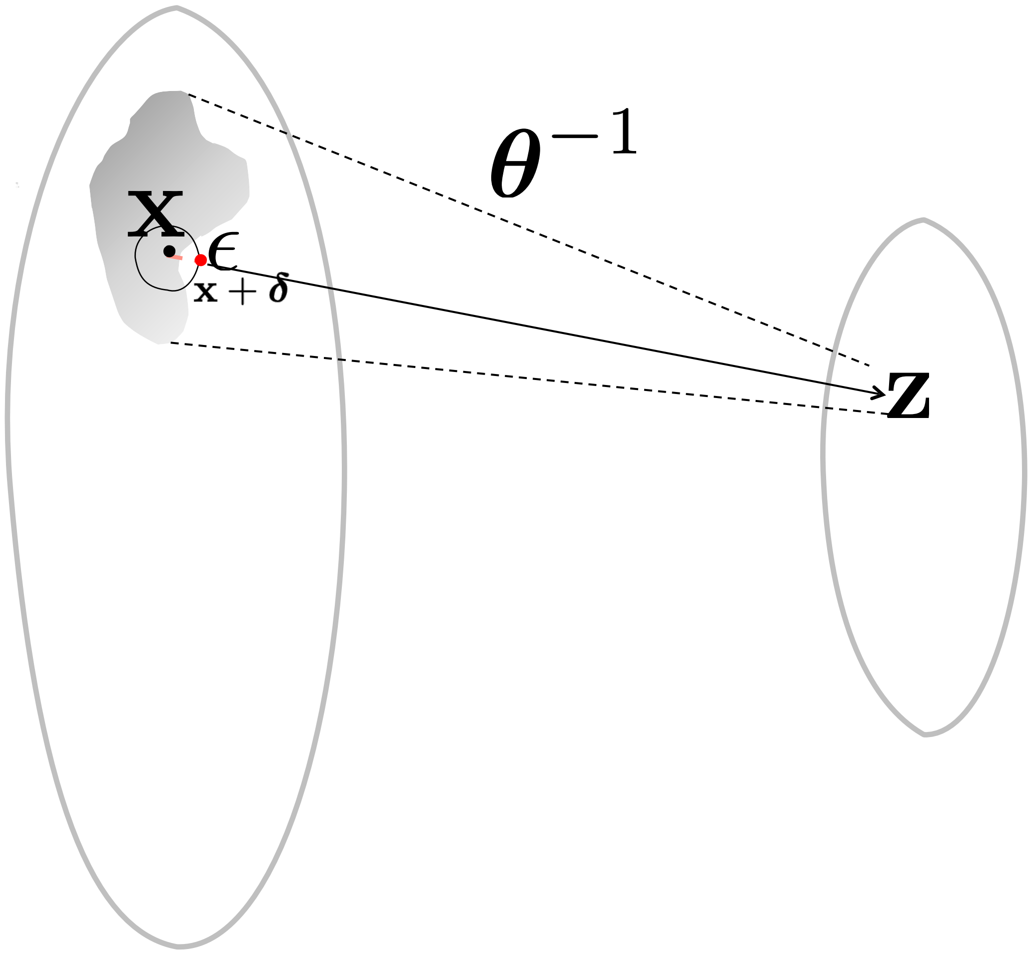


"adversarial training"

$$\boldsymbol{\theta}^{\star} = \arg\min_{\boldsymbol{\theta}} \ell\big(\boldsymbol{\theta}(\mathbf{x}+\boldsymbol{\delta}^{\star}), y\big) \quad \text{where}$$

$$\boldsymbol{\delta}^{\star} = \arg\max_{||\boldsymbol{\delta}||_p<\epsilon} \ell\big(\boldsymbol{\theta}\left(\mathbf{x}+\boldsymbol{\delta}\right), y\big)$$

Robust Model $p(z|\mathbf{x})$

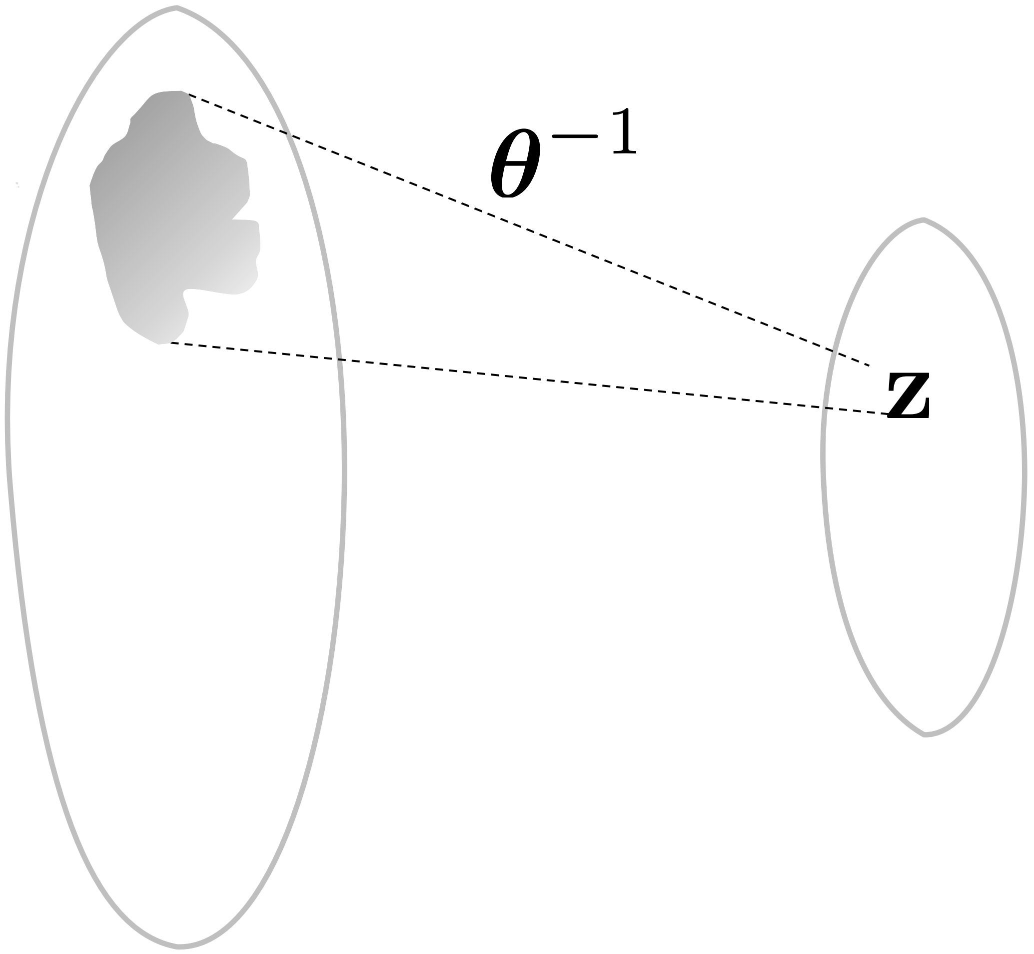


"adversarial training"

$$\boldsymbol{\theta}^{\star} = \arg\min_{\boldsymbol{\theta}} \ell\big(\boldsymbol{\theta}(\mathbf{x}+\boldsymbol{\delta}^{\star}), y\big) \quad \text{where}$$
$$\boldsymbol{\delta}^{\star} = \arg\max_{||\boldsymbol{\delta}||_p<\epsilon} \ell\big(\boldsymbol{\theta}\left(\mathbf{x}+\boldsymbol{\delta}\right), y\big)$$

\- Decreases the accuracy on natural data

Robust Model $p(z|\mathbf{x})$



"adversarial training"

$$\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}(\mathbf{x}+\boldsymbol{\delta}^\star), y) \quad \text{where}$$
$$\boldsymbol{\delta}^\star = \arg \max_{||\boldsymbol{\delta}||_p < \epsilon} \ell(\boldsymbol{\theta}(\mathbf{x}+\boldsymbol{\delta}), y)$$

- \- Decreases the accuracy on natural data
- \+ Develops "generative" behavior

Looking at the input gradients

Looking at the input gradients

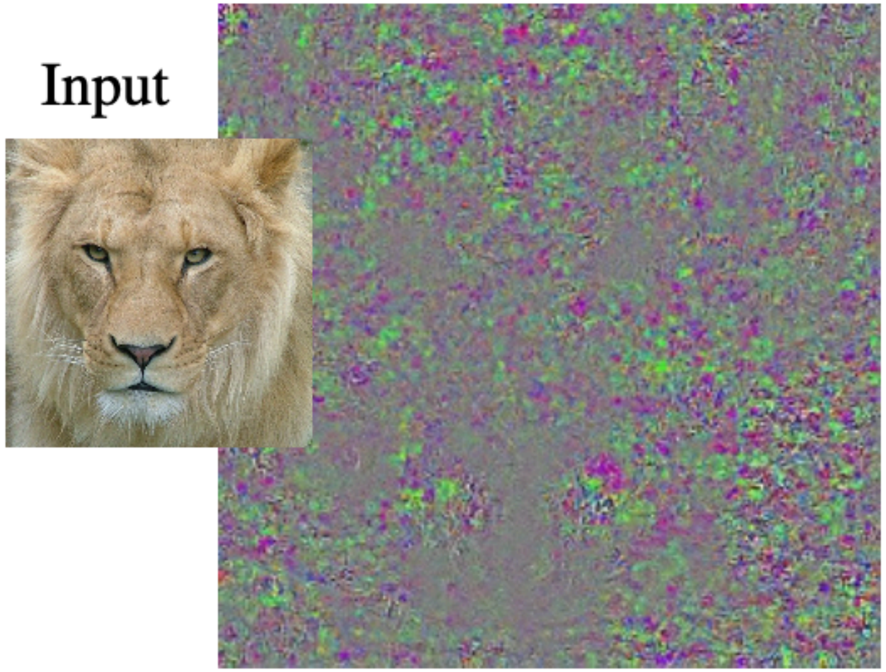
$$\nabla_{\mathbf{x}} \ell_{\mathrm{CE}}(\mathbf{x}, z; \boldsymbol{\theta})$$

Looking at the input gradients

$$\nabla_{\mathbf{x}} \ell_{\mathrm{CE}}(\mathbf{x}, z; \boldsymbol{\theta})$$

Standard, non-robust

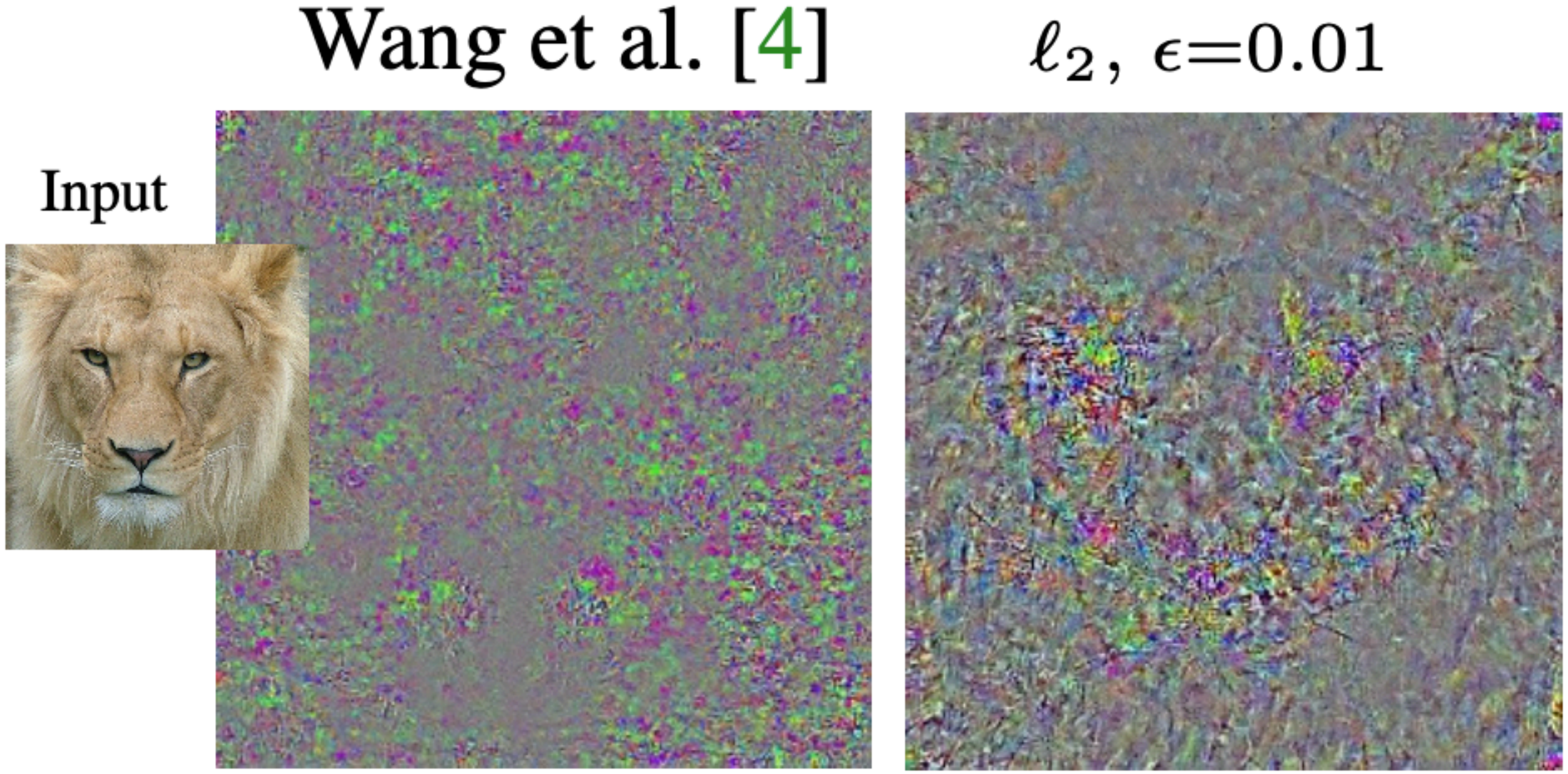
Wang et al. [4]



Looking at the input gradients

$$\nabla_{\mathbf{x}} \ell_{\text{CE}}(\mathbf{x}, z; \boldsymbol{\theta})$$

Standard, non-robust

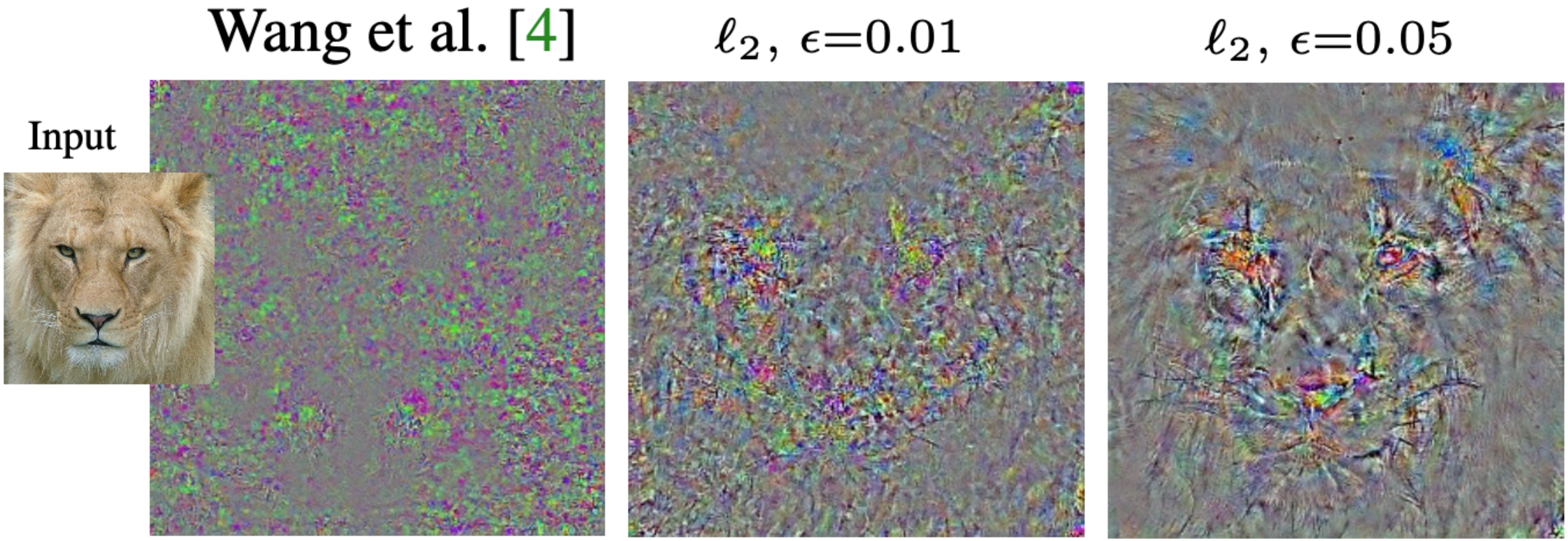


Looking at the input gradients

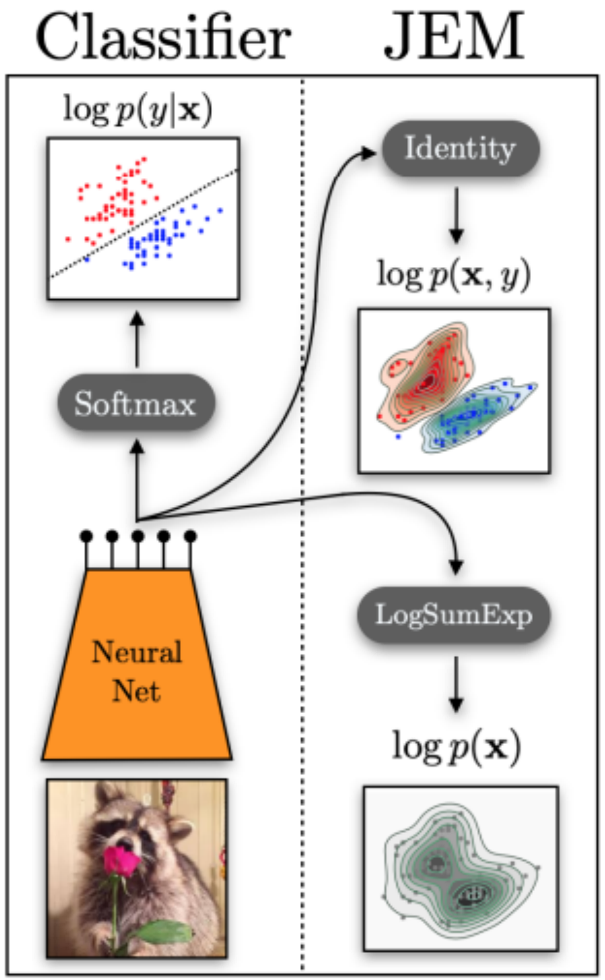
$$\nabla_{\mathbf{x}} \ell_{\mathrm{CE}}(\mathbf{x}, z; \boldsymbol{\theta})$$

Standard, non-robust

"Robust" family



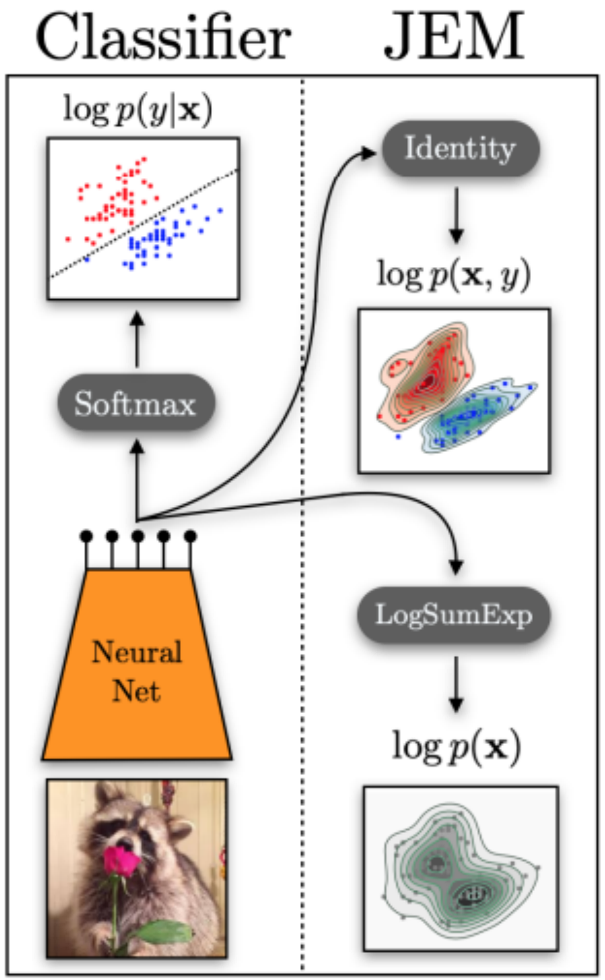
Why Robust Models
behave as Generative?



[image is from JEM - ICLR20]

[JEM - Your classifier is secretly an energy-based model, and you should treat it like one, ICLR20]

Why Robust Models behave as Generative?



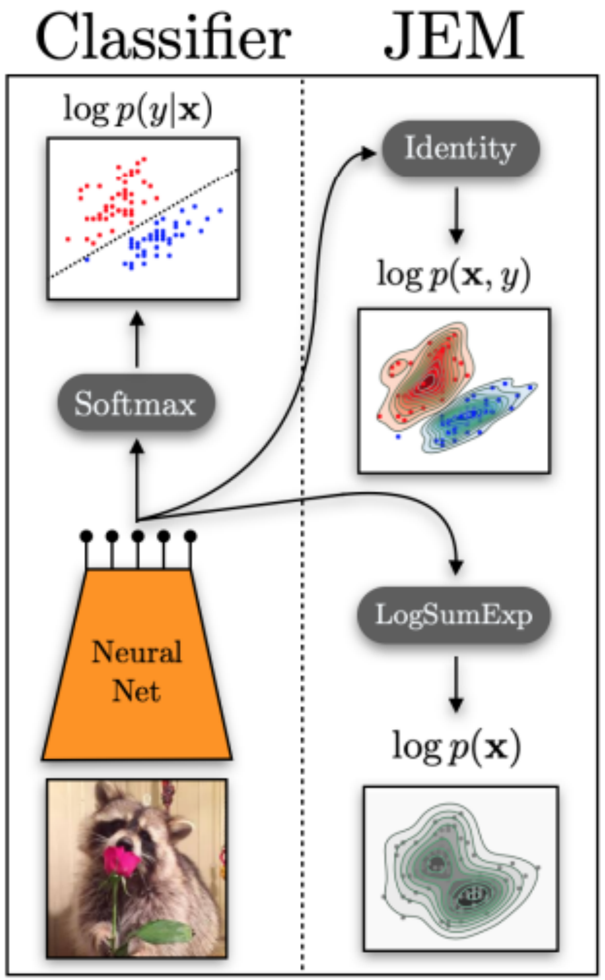
[image is from JEM - ICLR20]

“energy-based model”

$$p_\theta(\mathbf{x}) = \frac{\exp\big(-E_\theta(\mathbf{x})\big)}{P_\theta}$$

[JEM - Your classifier is secretly an energy-based model, and you should treat it like one, ICLR20]

Why Robust Models behave as Generative?



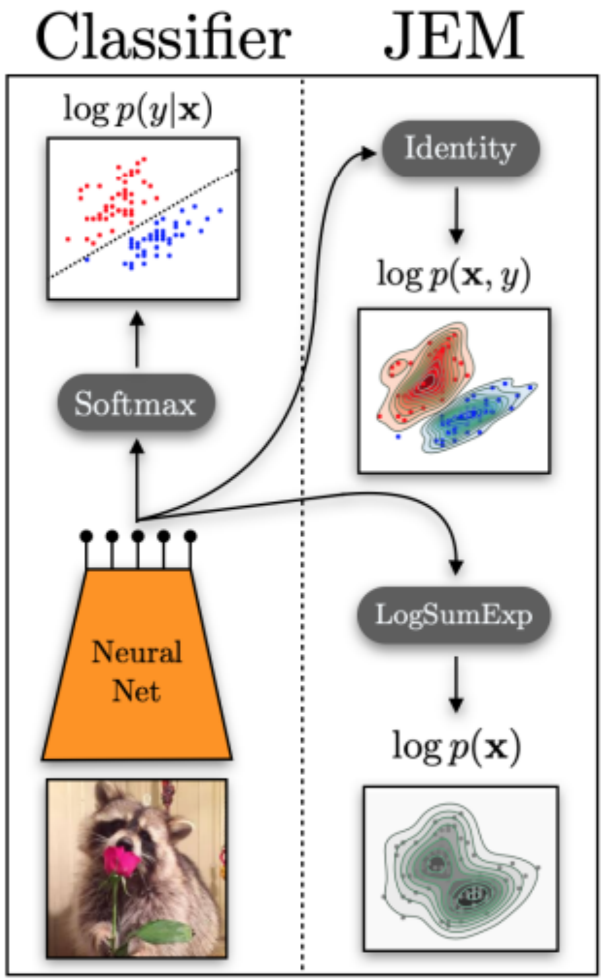
[image is from JEM - ICLR20]

"energy-based model"

$$p_\theta(\mathbf{x}) = \frac{\exp\big(-E_\theta(\mathbf{x})\big)}{P_\theta}$$

Classifier

$$p(z = i|\mathbf{x}) = \frac{\exp F_\theta(x)[i]}{\sum_{i=1}^{K} \exp F_\theta(x)[i]}$$



[image is from JEM - ICLR20]

"energy-based model"

$$p_\theta(\mathbf{x}) = \frac{\exp\big(-E_\theta(\mathbf{x})\big)}{P_\theta}$$

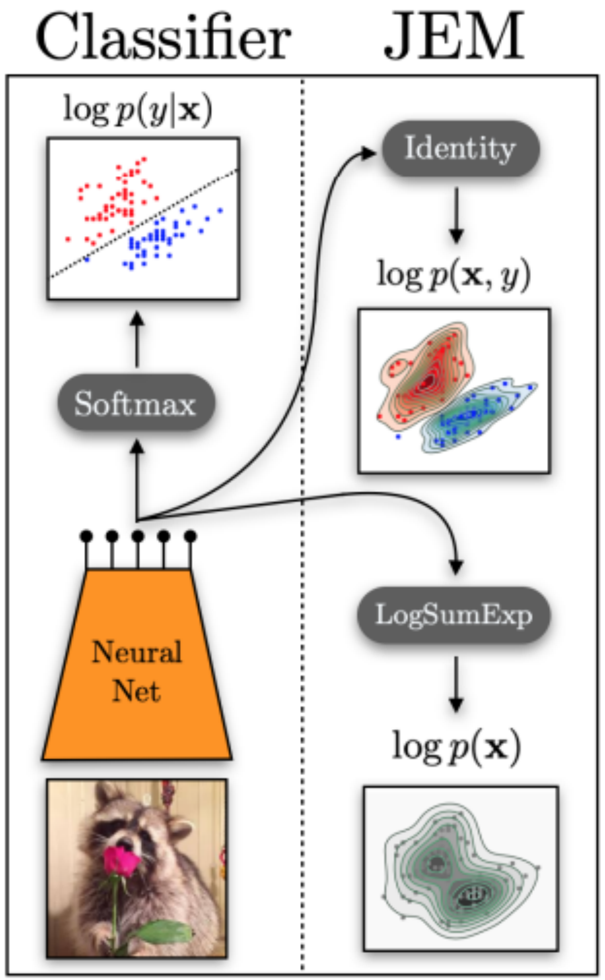
Classifier

$$p(z = i|\mathbf{x}) = \frac{\exp F_\theta(x)[i]}{\sum_{i=1}^{K} \exp F_\theta(x)[i]}$$

$$E_\theta(\mathbf{x}, z)$$

Joint energy: datum vs label

Why Robust Models
behave as Generative?



[image is from JEM - ICLR20]

"energy-based model"

$$p_\theta(\mathbf{x}) = \frac{\exp\left(-E_\theta(\mathbf{x})\right)}{P_\theta}$$

Classifier

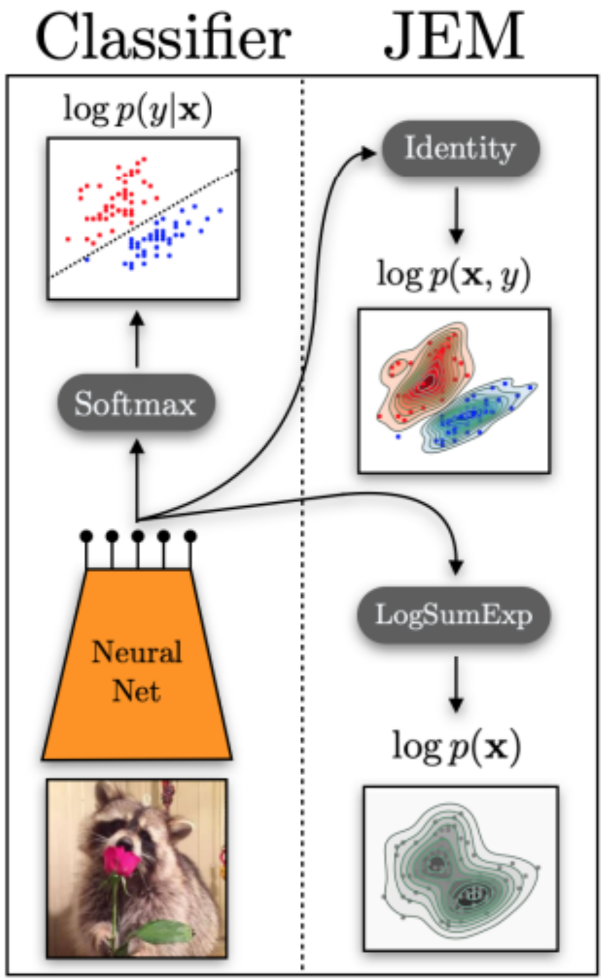
$$p(z=i|\mathbf{x}) = \frac{\exp F_\theta(x)[i]}{\sum_{i=1}^{K} \exp F_\theta(x)[i]}$$

$$E_\theta(\mathbf{x}, z) \qquad\qquad E_\theta(\mathbf{x})$$

Joint energy: datum vs label

Marginal: Energy of datum

Why Robust Models behave as Generative?



[image is from JEM - ICLR20]

"energy-based model"

$$p_\theta(\mathbf{x}) = \frac{\exp\big(-E_\theta(\mathbf{x})\big)}{P_\theta}$$

Classifier

$$p(z=i|\mathbf{x}) = \frac{\exp F_\theta(x)[i]}{\sum_{i=1}^{K} \exp F_\theta(x)[i]}$$

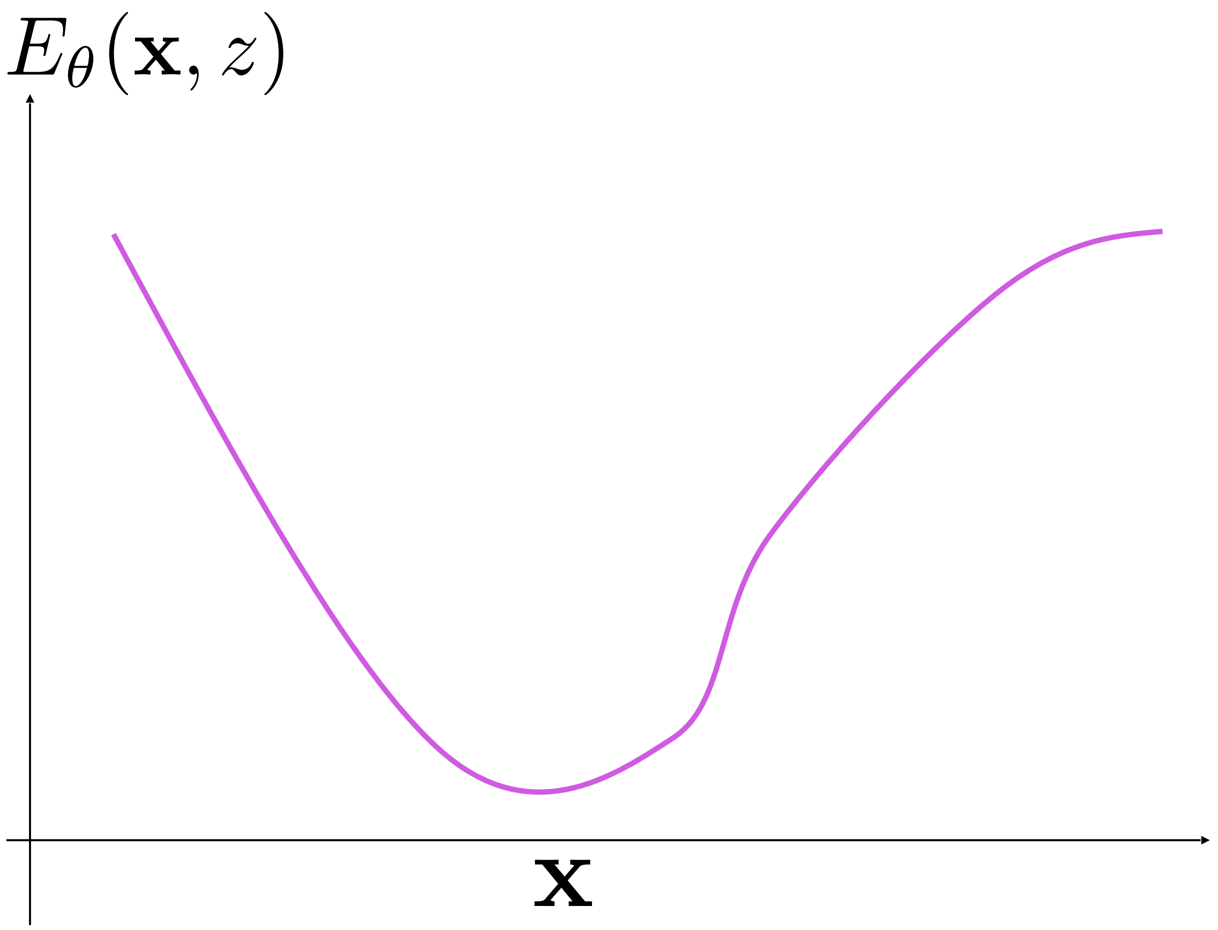
$$E_\theta(\mathbf{x}, z) \qquad E_\theta(\mathbf{x})$$

Joint energy: datum vs label

Marginal: Energy of datum

$$\ell_{\mathrm{CE}}(\mathbf{x}, z; \boldsymbol{\theta}) = E(\mathbf{x}, z; \boldsymbol{\theta}) - E(\mathbf{x}; \boldsymbol{\theta})$$

Why Robust Models behave as Generatives?

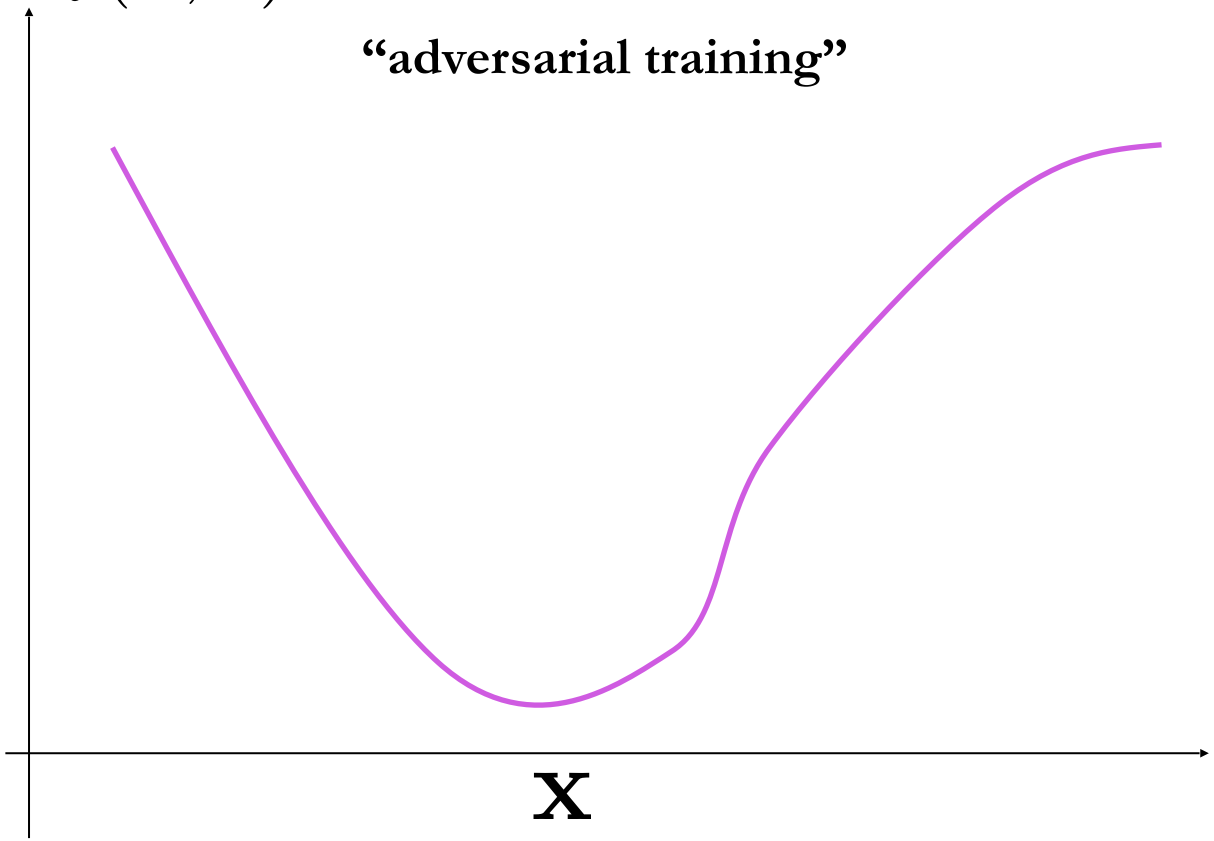


$$E_\theta(\mathbf{x}, z)$$

[Towards understanding generative capability of robust model , ICCV21]

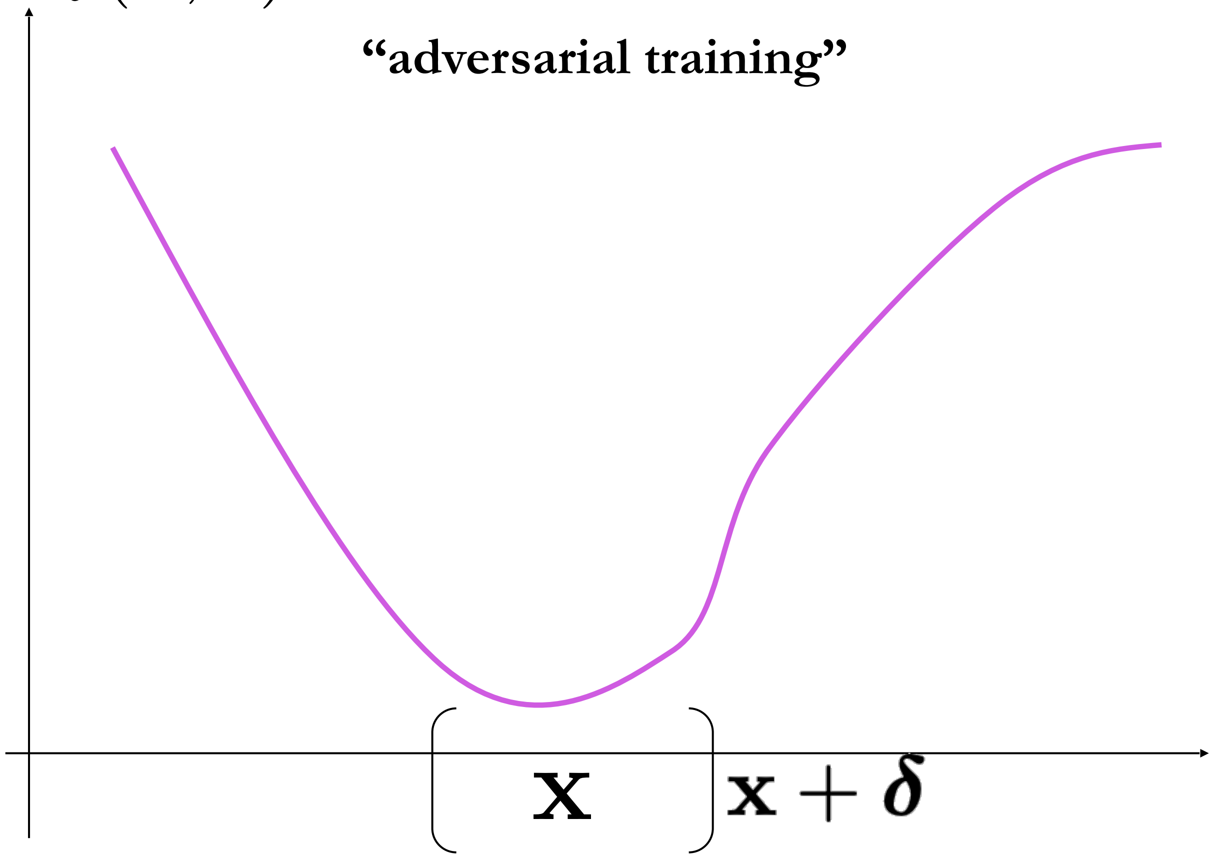
Why Robust Models behave as Generatives?

$E_\theta(\mathbf{x}, z)$



$E_\theta(\mathbf{x}, z)$

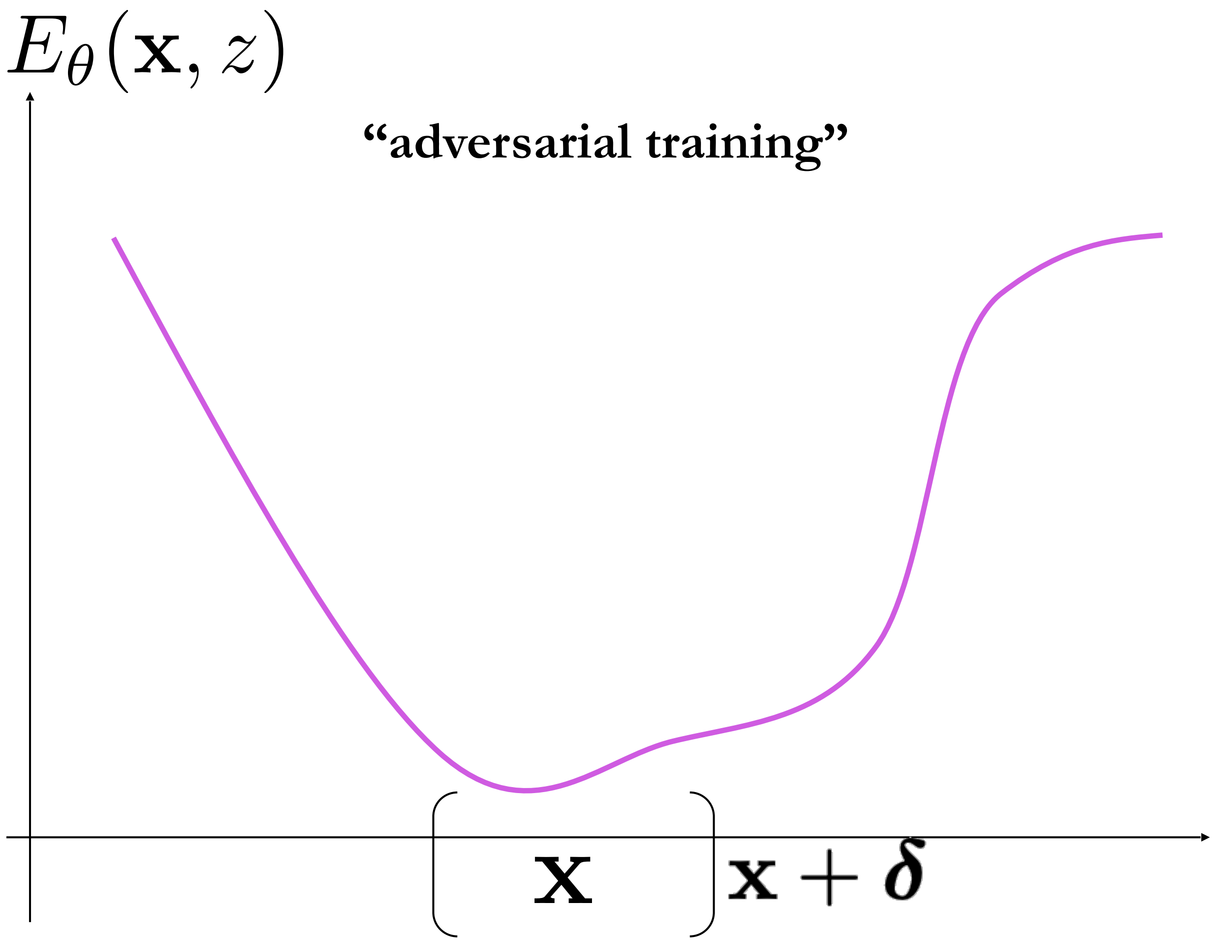
$E_\theta(\mathbf{x}, z)$



$E_\theta(\mathbf{x}, z)$

[Towards understanding generative capability of robust model , ICCV21]

Why Robust Models behave as Generatives?

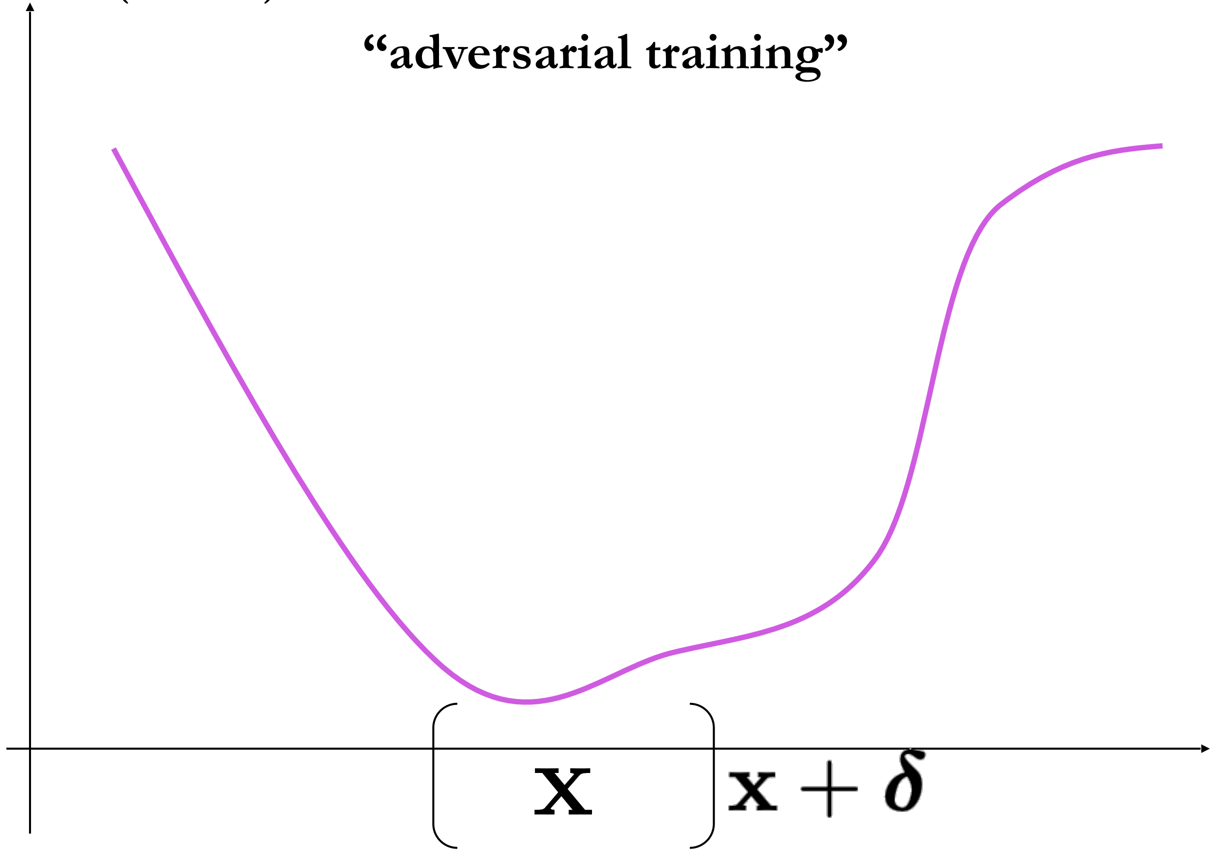


$E_\theta(\mathbf{x}, z)$

Why Robust Models behave as Generatives?

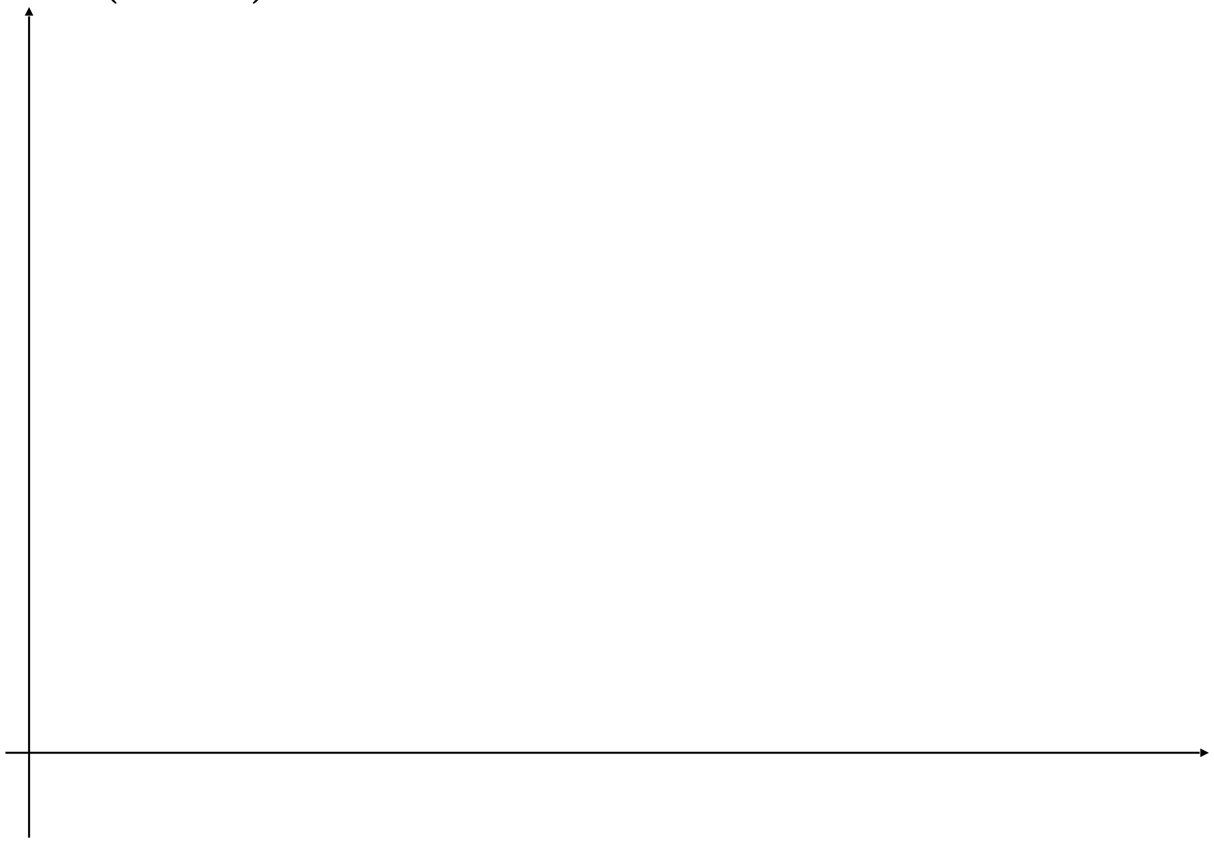
$E_\theta(\mathbf{x}, z)$

"adversarial training"



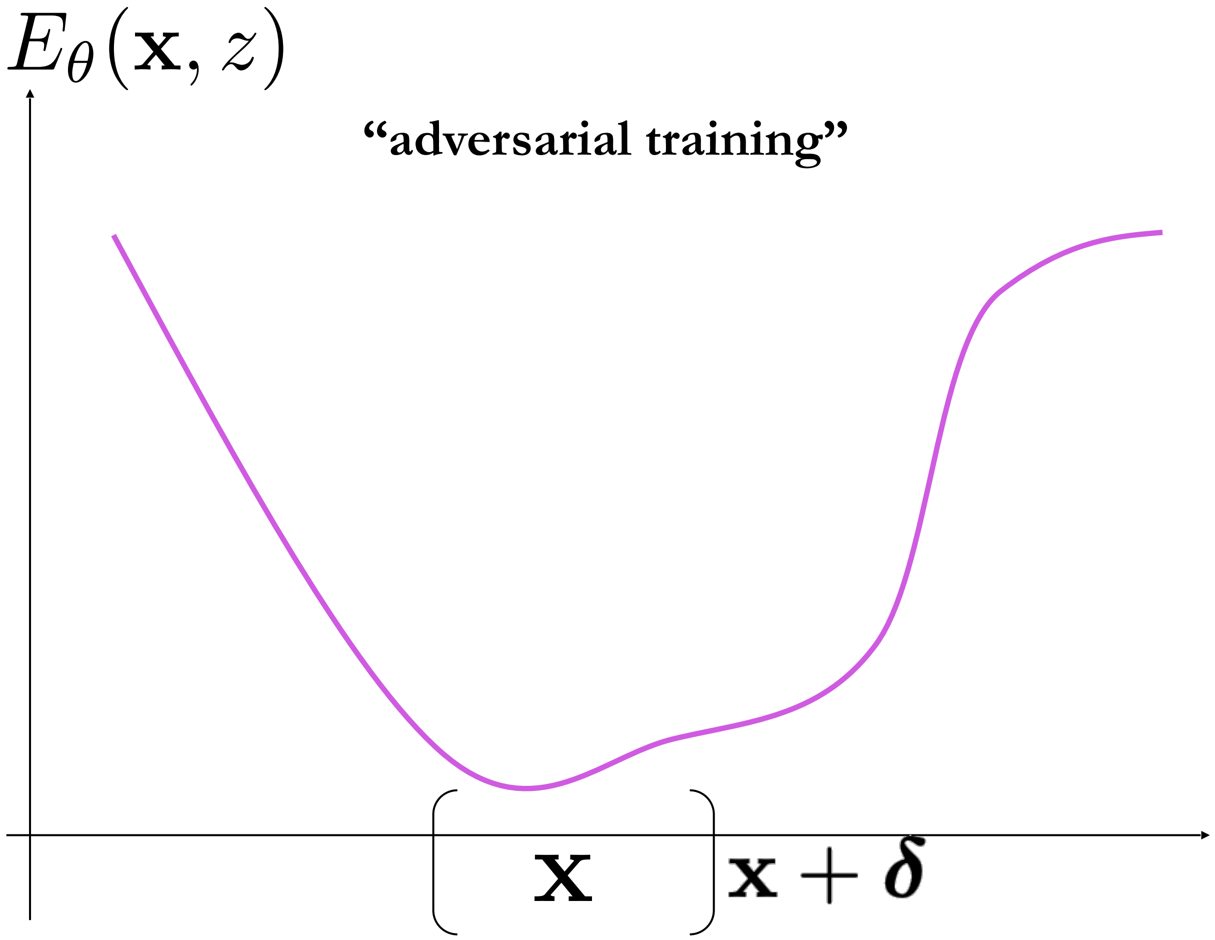
$E_\theta(\mathbf{x}, z)$

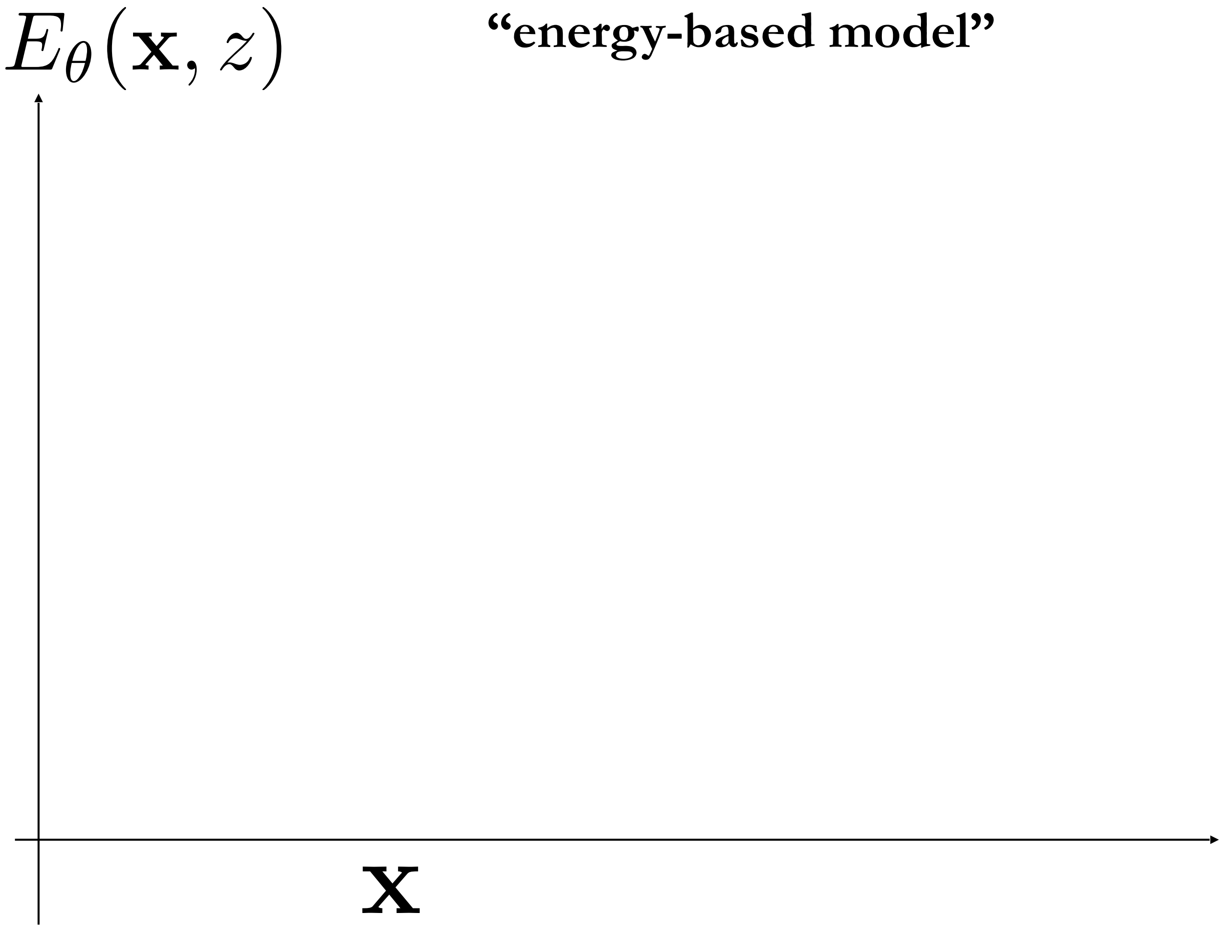
"energy-based model"



[Towards understanding generative capability of robust model , ICCV21]

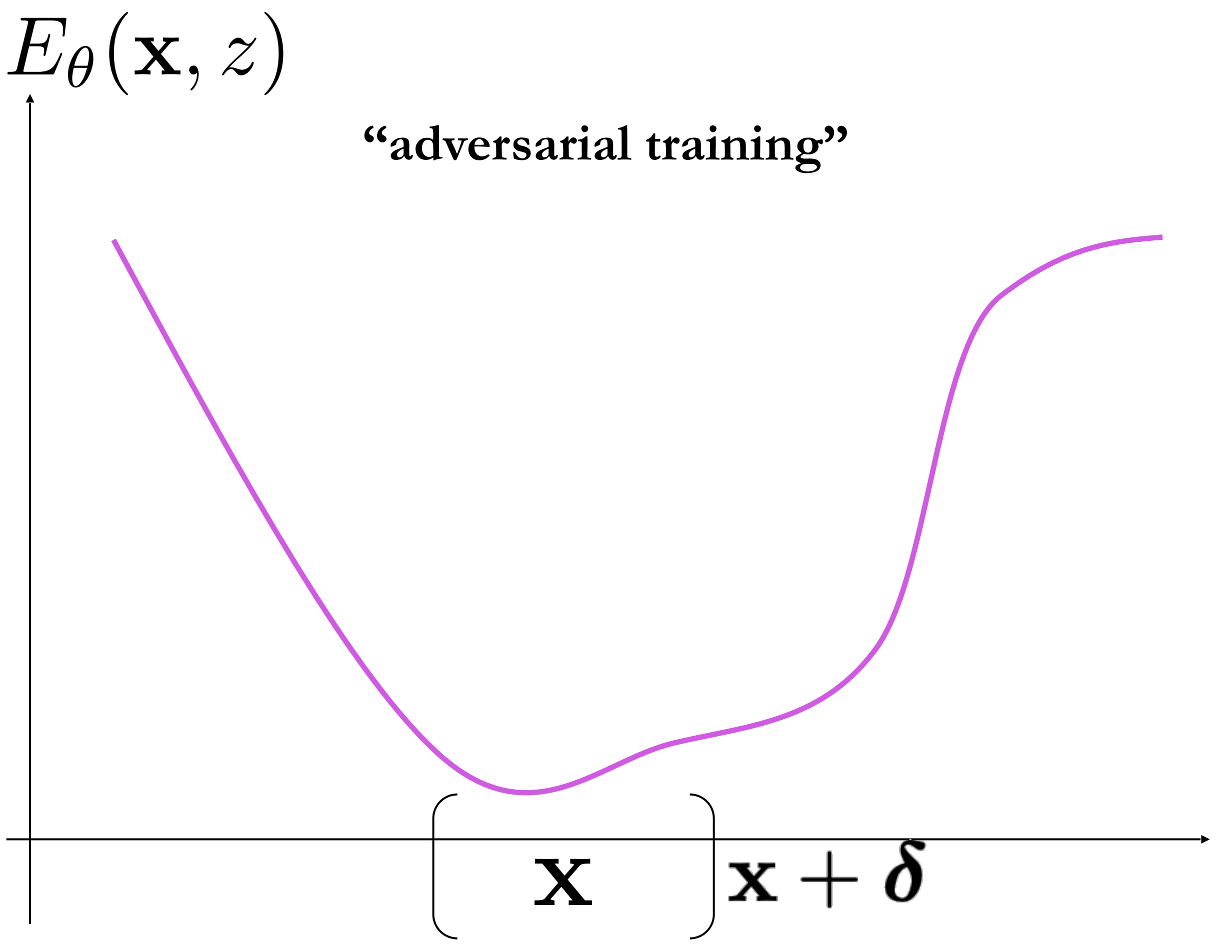
Why Robust Models behave as Generatives?

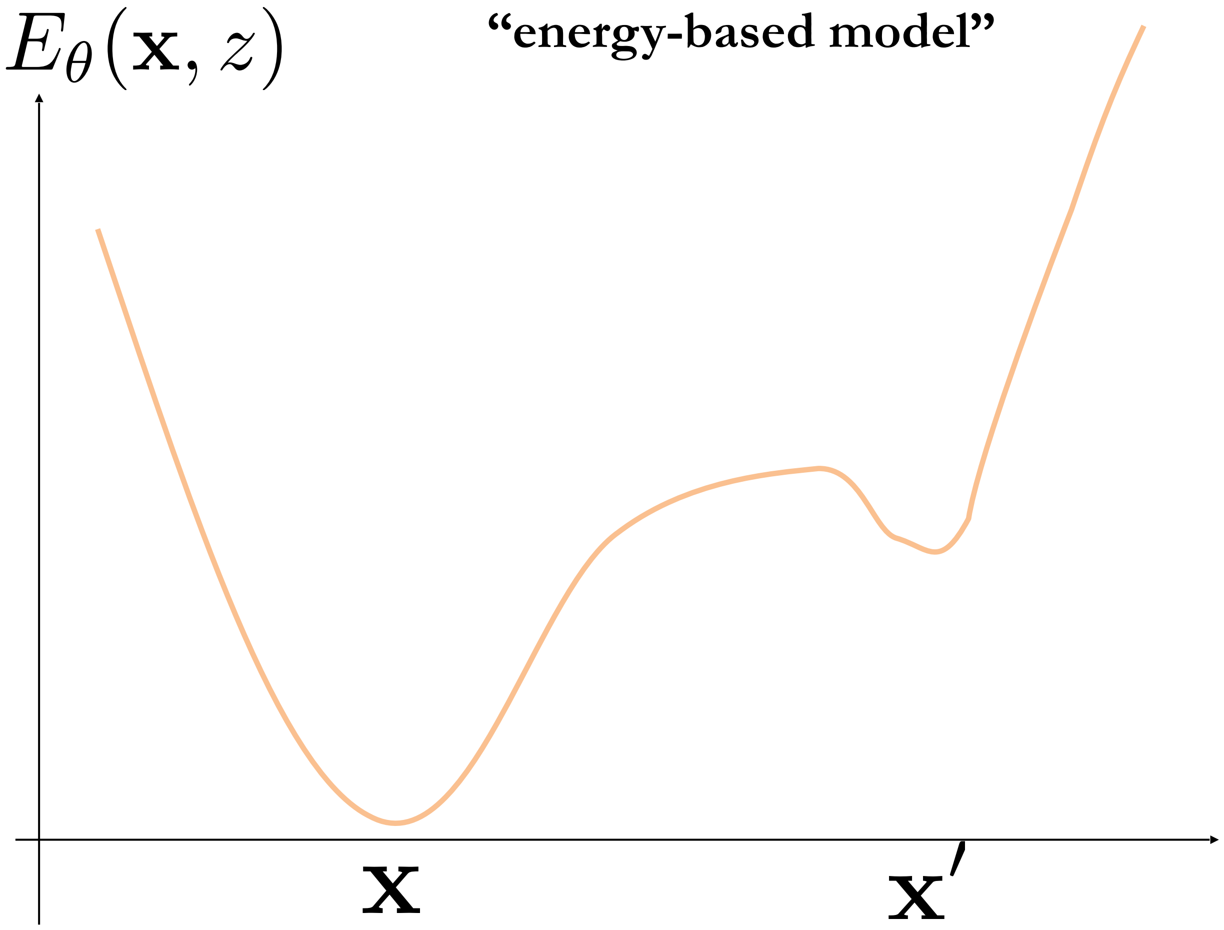




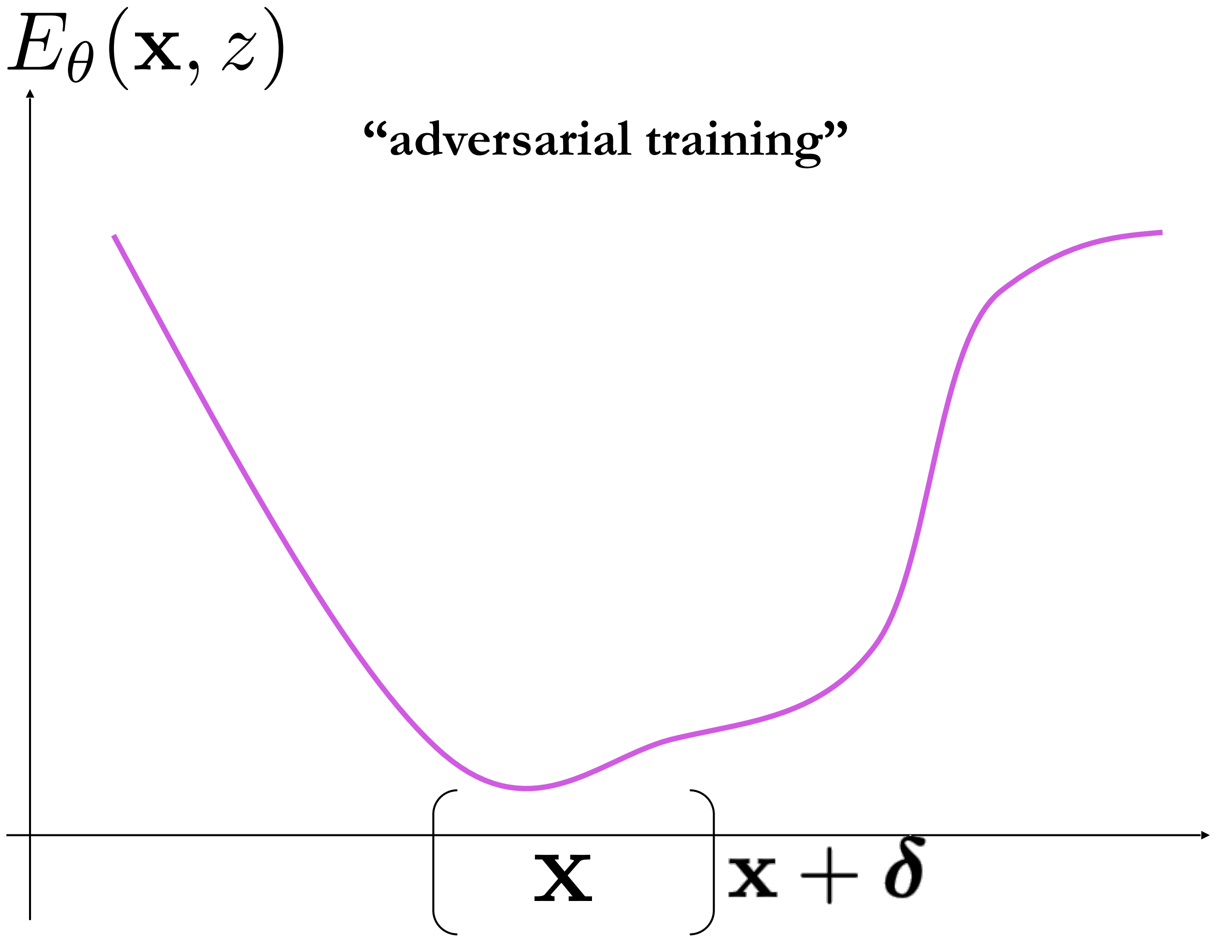
[Towards understanding generative capability of robust model , ICCV21]

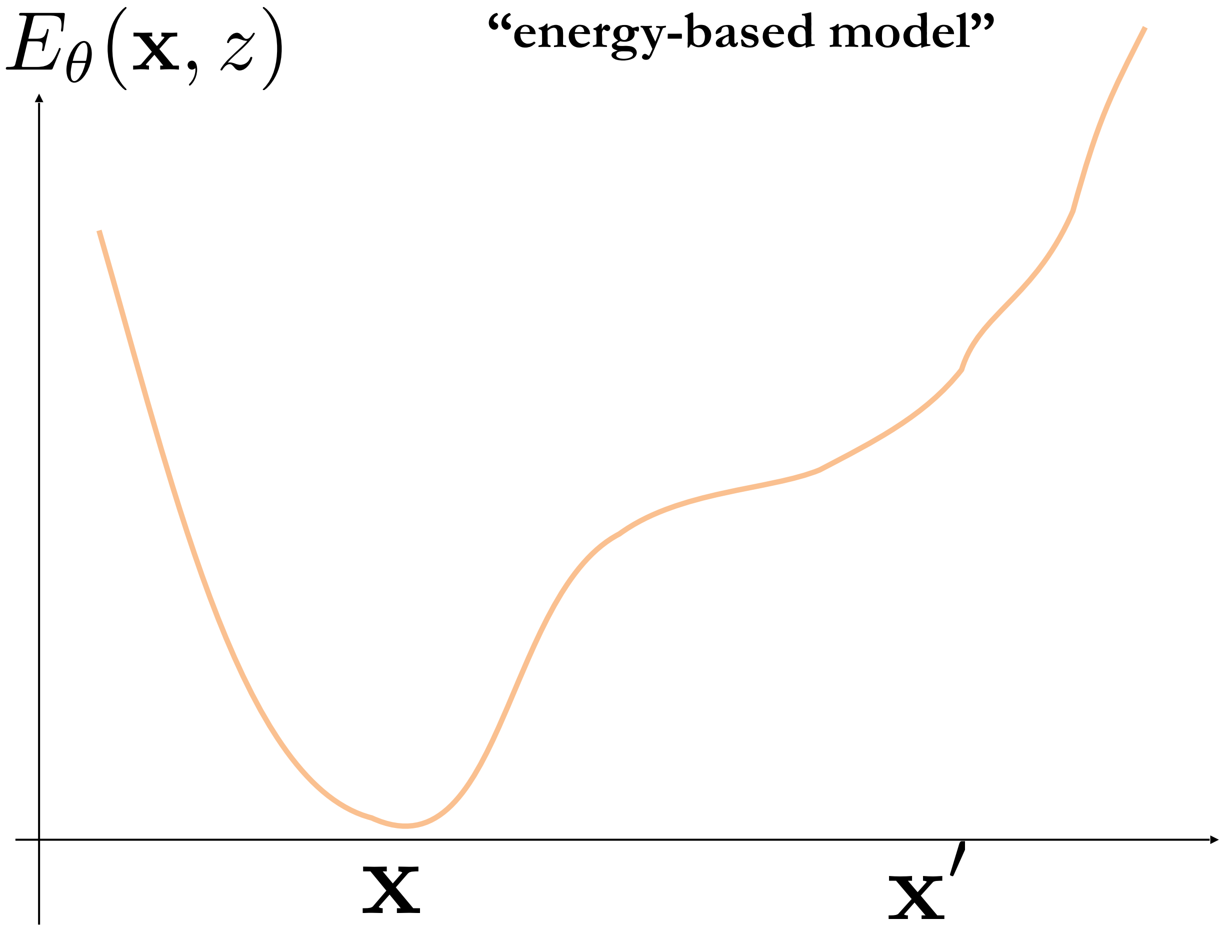
Why Robust Models
behave as Generatives?





[Towards understanding generative capability of robust model , ICCV21]





[Towards understanding generative capability of robust model , ICCV21]

Energy $E_\theta(\mathbf{x})$ vs attack strength

Energy $E_\theta(\mathbf{x})$ vs attack strength

Finding 1: Untargeted attacks* decrease $E_\theta(\mathbf{x})$ thus increase $p_\theta(\mathbf{x})$

* Untargeted attack = PGD (Projected Gradient Descent) attack

Energy $E_\theta(\mathbf{x})$ vs attack strength

Finding 1: Untargeted attacks* decrease $E_\theta(\mathbf{x})$ thus increase $p_\theta(\mathbf{x})$

In other words, untargeted attacks finds points with:

- High energy $E_\theta(\mathbf{x}, z)$ thus low $p_\theta(\mathbf{x}, z)$

Fool the classifier (known)

* Untargeted attack = PGD (Projected Gradient Descent) attack

Energy $E_\theta(\mathbf{x})$ vs attack strength

Finding 1: Untargeted attacks* decrease $E_\theta(\mathbf{x})$ thus increase $p_\theta(\mathbf{x})$

In other words, untargeted attacks finds points with:

- High energy $E_\theta(\mathbf{x}, z)$ thus low $p_\theta(\mathbf{x}, z)$
- Low $E_\theta(\mathbf{x})$, highly likely for the model $p_\theta(\mathbf{x})$

Fool the classifier (known)

New adversarial points are more likely to exist than the natural data points! (less known)

* Untargeted attack = PGD (Projected Gradient Descent) attack

Energy $E_\theta(\mathbf{x})$ vs attack strength

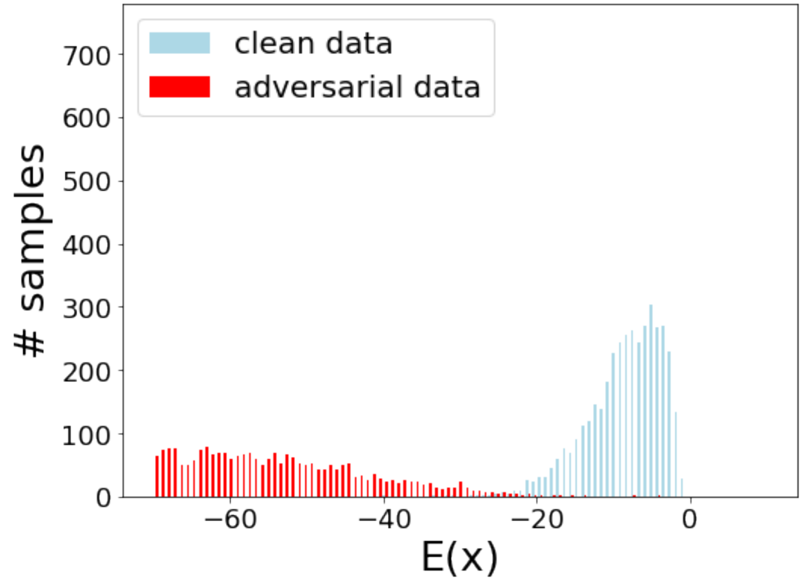
Finding 1: Untargeted attacks* decrease $E_\theta(\mathbf{x})$ thus increase $p_\theta(\mathbf{x})$

In other words, untargeted attacks finds points with:

- High energy $E_\theta(\mathbf{x}, z)$ thus low $p_\theta(\mathbf{x}, z)$
- Low $E_\theta(\mathbf{x})$, highly likely for the model $p_\theta(\mathbf{x})$

Fool the classifier (known)

New adversarial points are more likely to exist than the natural data points! (less known)



* Untargeted attack = PGD (Projected Gradient Descent) attack

Energy $E_\theta(\mathbf{x})$ vs attack strength

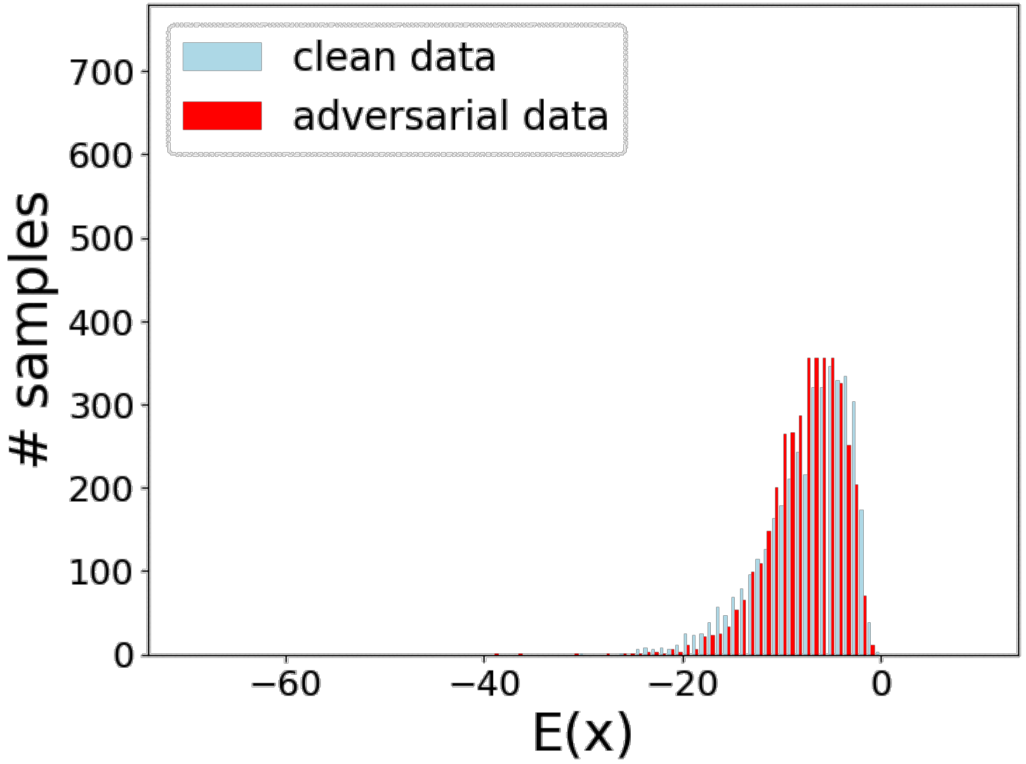
Energy $E_\theta(\mathbf{x})$ vs attack strength

Finding 2: $E_\theta(\mathbf{x})$ decreases as the attack "strength" increases

* attack strength= iterations in PGD

Energy $E_\theta(\mathbf{x})$ vs attack strength

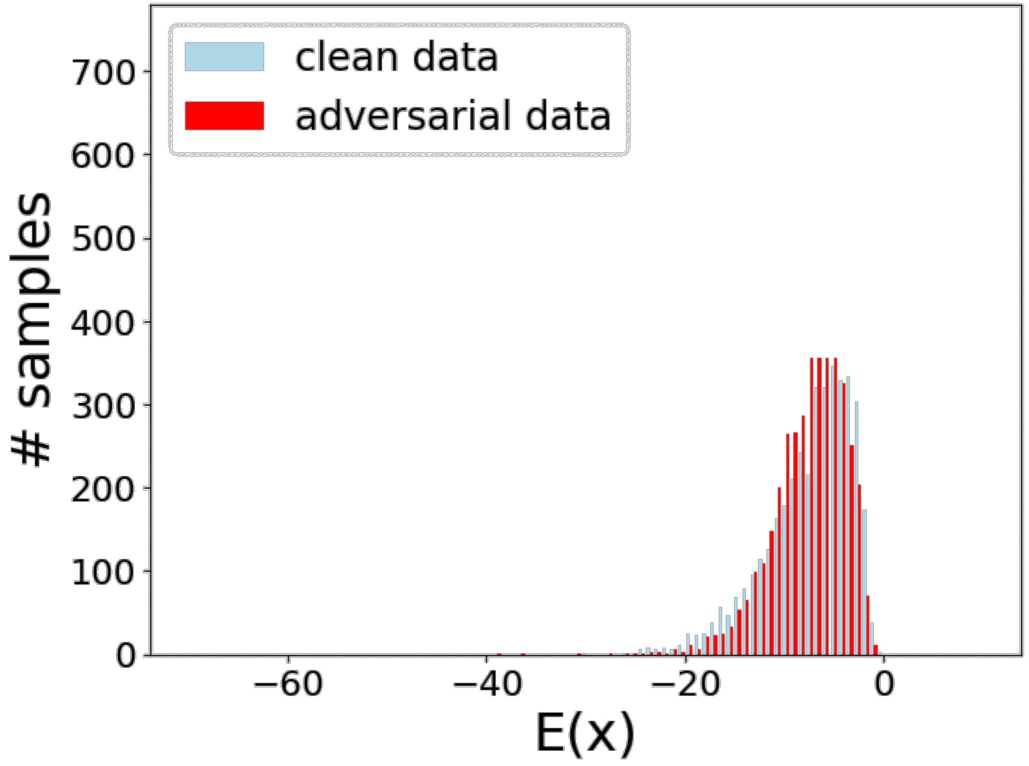
Finding 2: $E_\theta(\mathbf{x})$ decreases as the attack "strength" increases

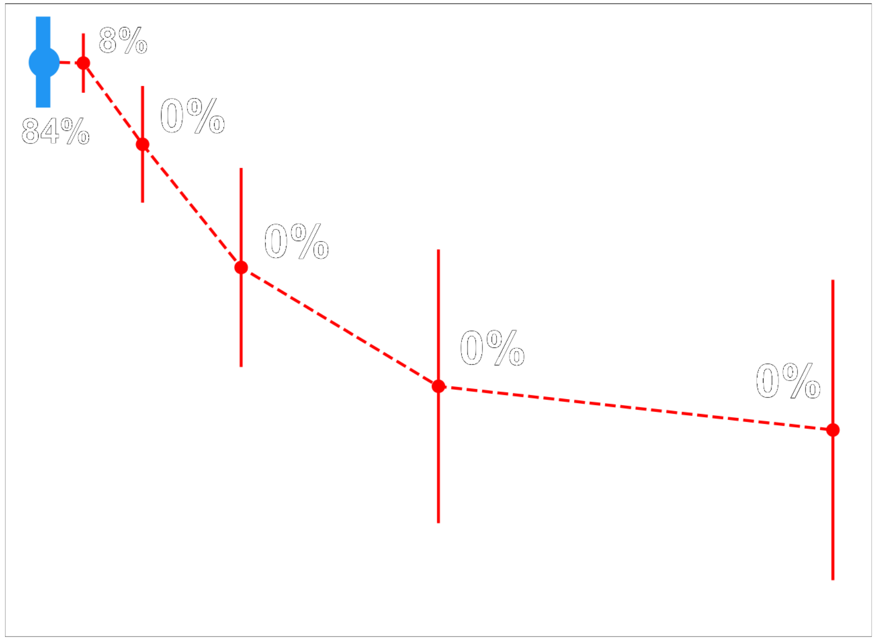


[*] attack strength= iterations in PGD

Energy $E_\theta(\mathbf{x})$ vs attack strength

Finding 2: $E_\theta(\mathbf{x})$ decreases as the attack "strength" increases

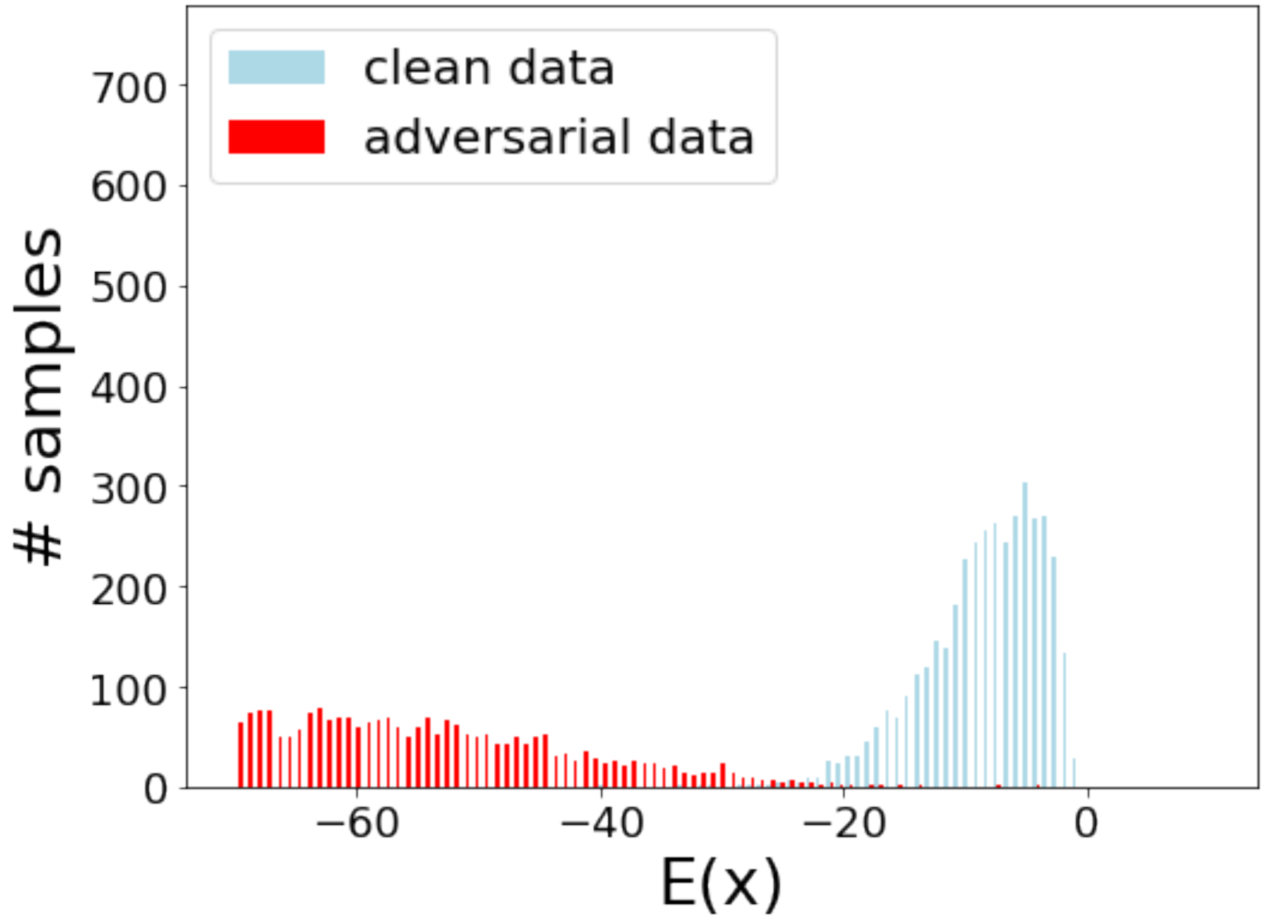




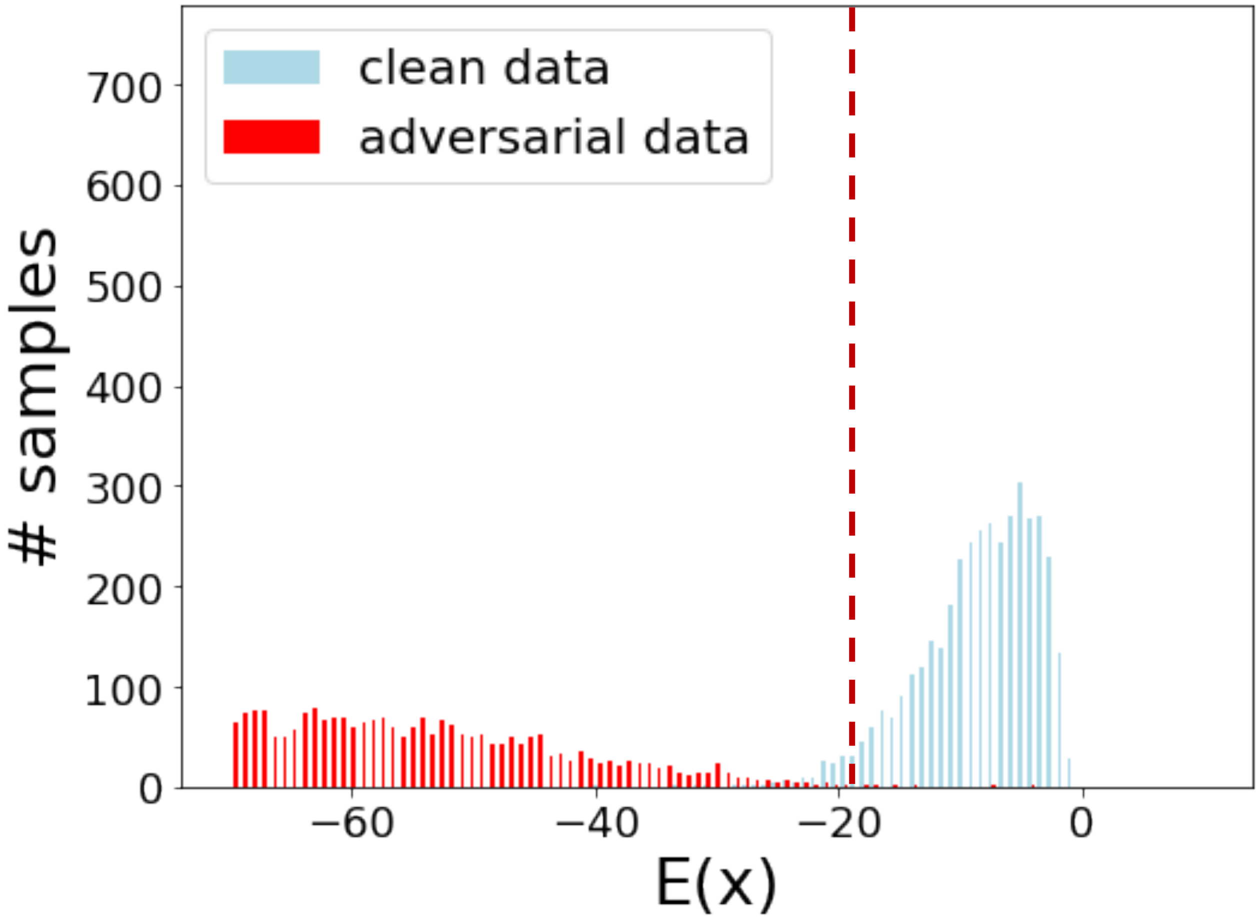
* attack strength= iterations in PGD

Efficient and Effective Attack Detector

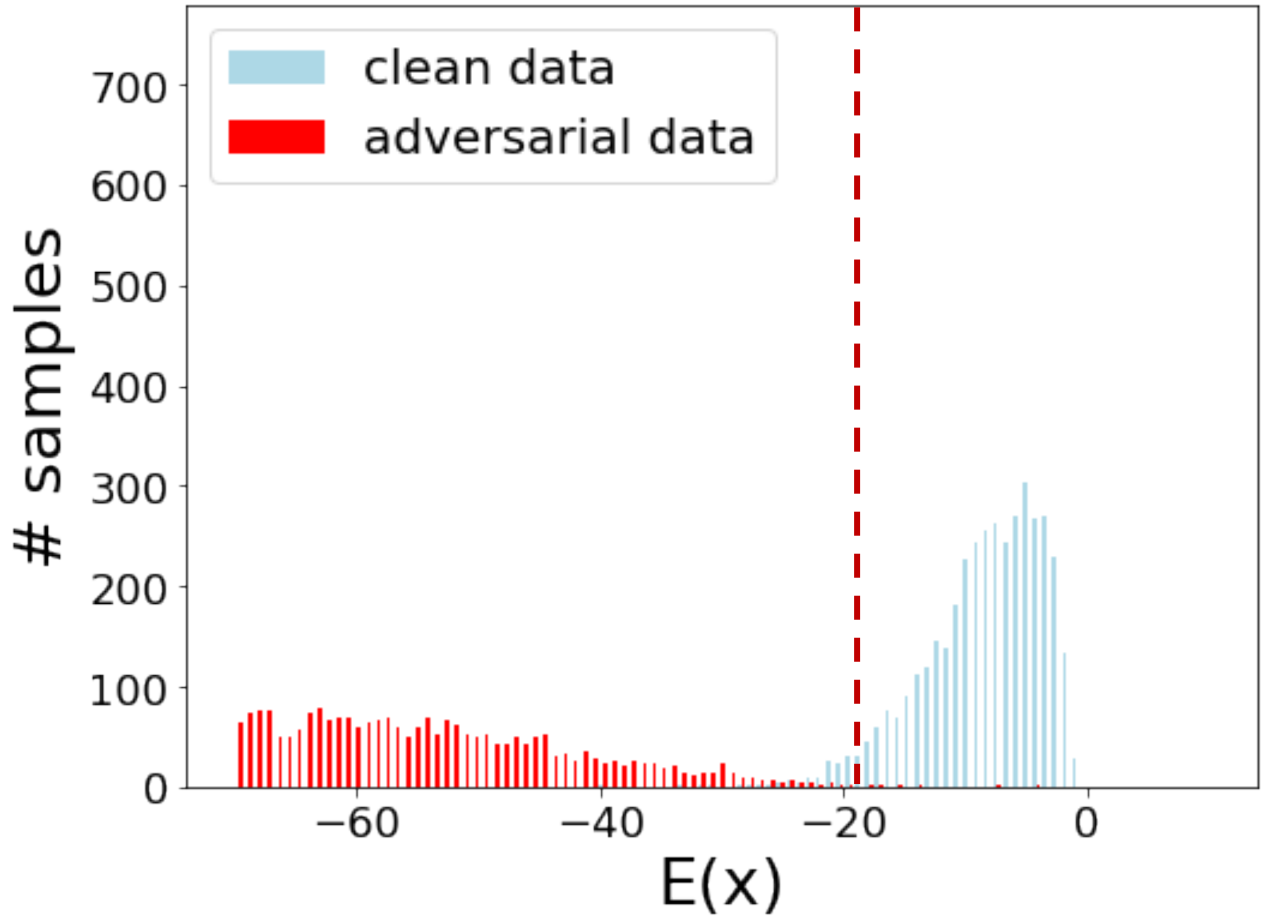
Efficient and Effective Attack Detector



Efficient and Effective Attack Detector



Efficient and Effective Attack Detector



Efficient and Effective Attack Detector

Efficient and Effective Attack Detector

Dataset	Defense	Attack	DR	FPR
imagenette [8]	**Energy** (ResNet10)	PGD (8)	98.24	1.37
		PGD (16)	99.6	0.00

Efficient and Effective Attack Detector

Dataset	Defense	Attack	DR	FPR
imagenette [8]	**Energy** (ResNet10)	PGD (8)	98.24	1.37
		PGD (16)	99.6	0.00

Efficient and Effective Attack Detector

Dataset	Defense	Attack	DR	FPR
imagenette [8]	**Energy** (ResNet10)	PGD (8)	98.24	1.37
		PGD (16)	99.6	0.00
CIFAR-10 [17]	**Energy** (ResNet10)	PGD (8)	98.38	1.62
		APGD (8)	85.45	1.19
	KD+BU [9]	PGD (8)	92.27	0.96
	LID [21]	PGD (8)	94.39	1.81

Efficient and Effective Attack Detector

Dataset	Defense	Attack	DR	FPR
imagenette [8]	**Energy** (ResNet10)	PGD (8)	98.24	1.37
		PGD (16)	99.6	0.00
CIFAR-10 [17]	**Energy** (ResNet10)	PGD (8)	98.38	1.62
		APGD (8)	85.45	1.19
	KD+BU [9]	PGD (8)	92.27	0.96
	LID [21]	PGD (8)	94.39	1.81

Efficient and Effective Attack Detector

Dataset	Defense	Attack	DR	FPR
imagenette [8]	**Energy** (ResNet10)	PGD (8)	98.24	1.37
		PGD (16)	99.6	0.00
CIFAR-10 [17]	**Energy** (ResNet10)	PGD (8)	98.38	1.62
		APGD (8)	85.45	1.19
	KD+BU [9]	PGD (8)	92.27	0.96
	LID [21]	PGD (8)	94.39	1.81

Detector may suffer from: (1) targeted attacks (2) AutoAttack

Efficient and Effective Attack Detector

Can we bypass the detector?

High-Energy PGD

High-Energy PGD

Fool the classifier yet keep energy like natural data

High-Energy PGD

Fool the classifier yet keep energy like natural data

$$\arg\max_{\boldsymbol{\delta}} \left[\mathcal{L}\Big(\boldsymbol{\theta}(\mathbf{x} + \boldsymbol{\delta}), y \Big) + \lambda \, E_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta}) \right]$$

High-Energy PGD

Fool the classifier yet keep energy like natural data

High-Energy PGD

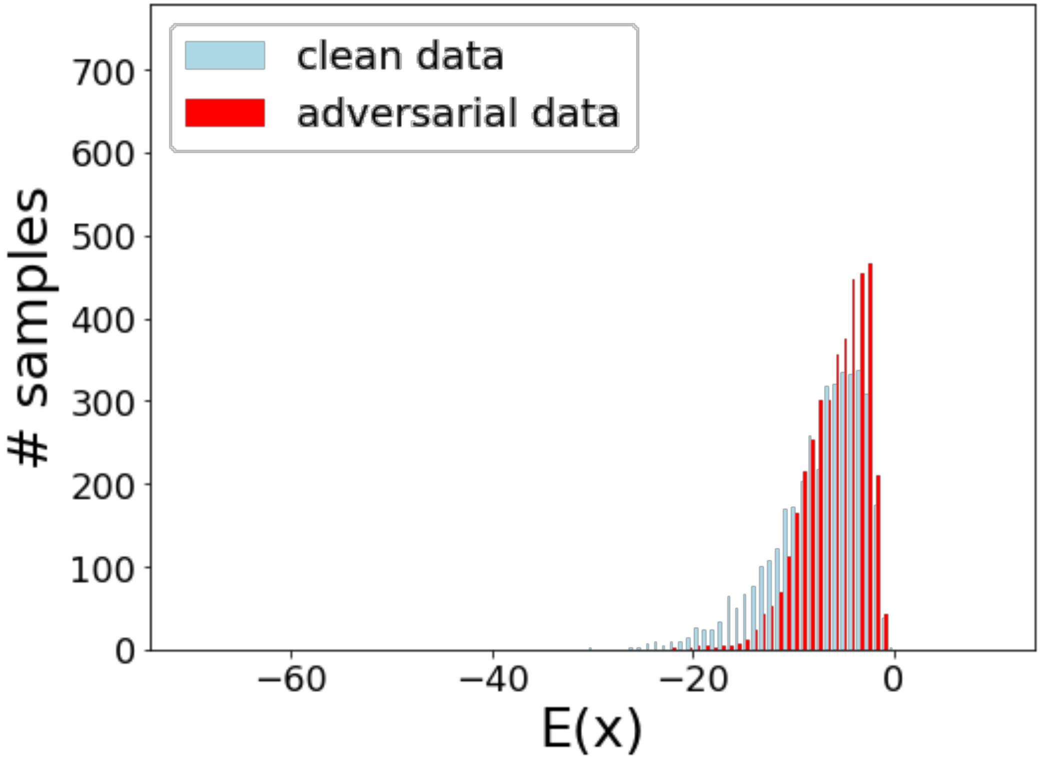
Fool the classifier yet keep energy like natural data

$$\mathbf{x}^* = \text{clip}_\epsilon\left[\mathbf{x}^* + \alpha\,\text{sign}\left[\nabla_{\mathbf{x}^*}\mathcal{L}(\boldsymbol{\theta}(\mathbf{x}^*), y) + \lambda E_{\boldsymbol{\theta}}(\mathbf{x}^*)\right]\right]$$

High-Energy PGD

Fool the classifier yet keep energy like natural data

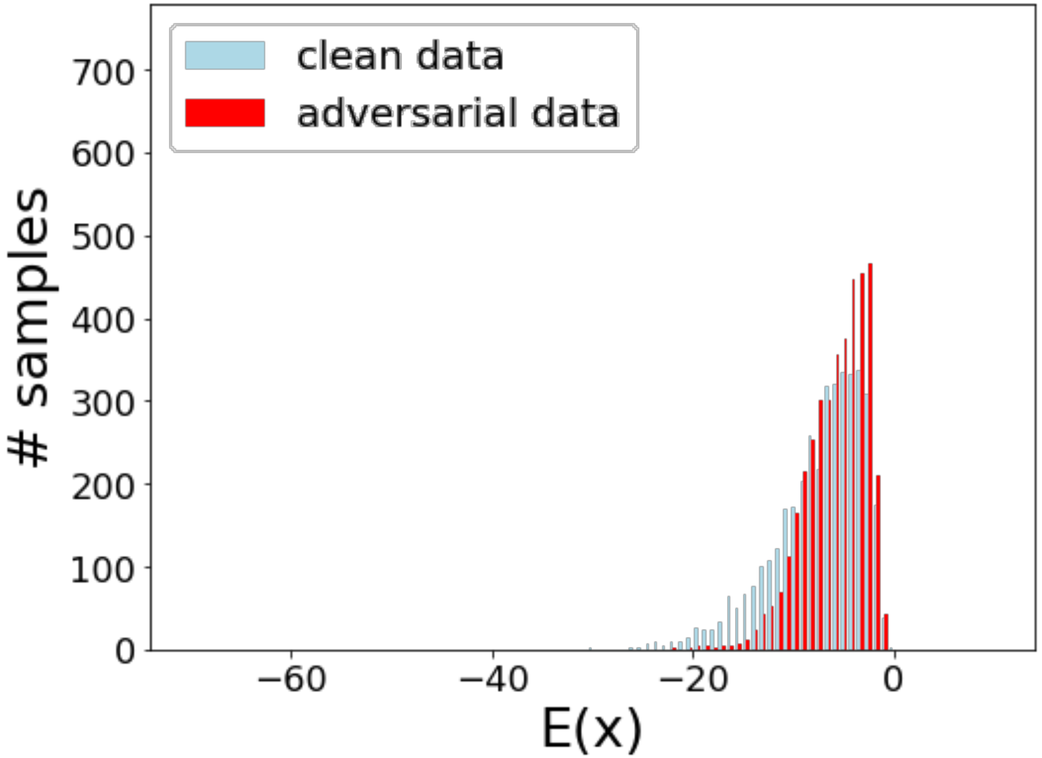
$$\mathbf{x}^* = \text{clip}_\epsilon\left[\mathbf{x}^* + \alpha\,\text{sign}\left[\nabla_{\mathbf{x}^*}\mathcal{L}(\boldsymbol{\theta}(\mathbf{x}^*), y) + \lambda E_{\boldsymbol{\theta}}(\mathbf{x}^*)\right]\right]$$

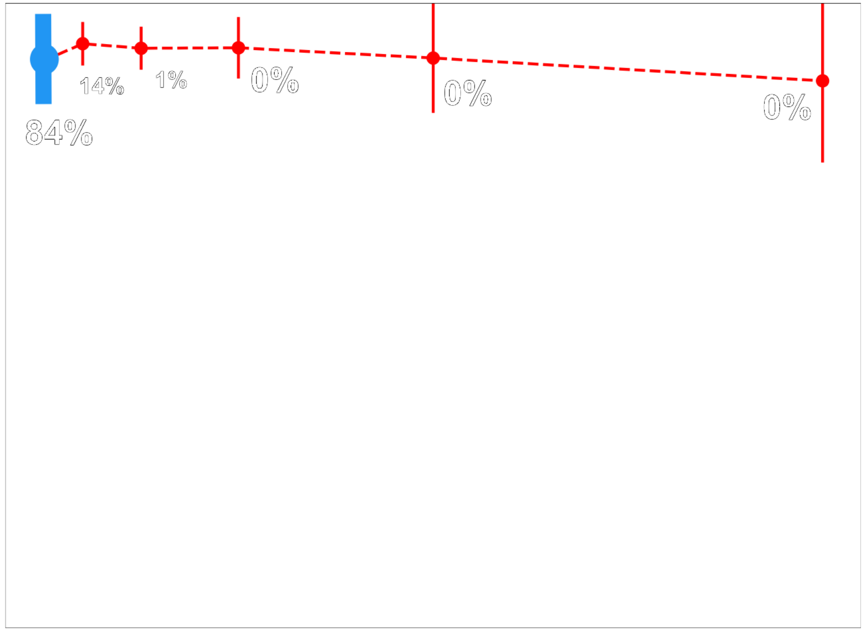


High-Energy PGD

Fool the classifier yet keep energy like natural data

$$\mathbf{x}^* = \text{clip}_\epsilon \left[\mathbf{x}^* + \alpha\, \text{sign}\left[\nabla_{\mathbf{x}^*}\mathcal{L}(\boldsymbol{\theta}(\mathbf{x}^*), y) + \lambda E_{\boldsymbol{\theta}}(\mathbf{x}^*)\right]\right]$$





Exploring the Connection between Robust and Generative Models

MAsk-Guided Image Synthesis by Inverting a Quasi-Robust Classifier [AAAI23]

Joint work: Mozhdeh Rouhsedaghat (USC) Masoud Monajatipoor (UCLA) C.-C. Jay Kuo (USC)

Method

$$\mathbf{x} \xrightarrow{\text{segment}} \mathbf{y}$$

Method

$$
\begin{tikzcd}
\mathbf{x} \arrow[r, dashed, "\text{segment}"] \arrow[d, "\text{encode}"] & \mathbf{y} \\
\mathbf{z} \arrow[d, "\text{invert}"] & \\
\mathbf{x}' &
\end{tikzcd}
$$

Method

$$
\begin{tikzcd}
\mathbf{x} \arrow[r, dashed, "\text{segment}"] \arrow[d, "\text{encode}"] \arrow[dd, bend right, dash] & \mathbf{y} \\
\mathbf{z} \arrow[d, "\text{invert}"] & \\
\mathbf{x}' &
\end{tikzcd}
$$

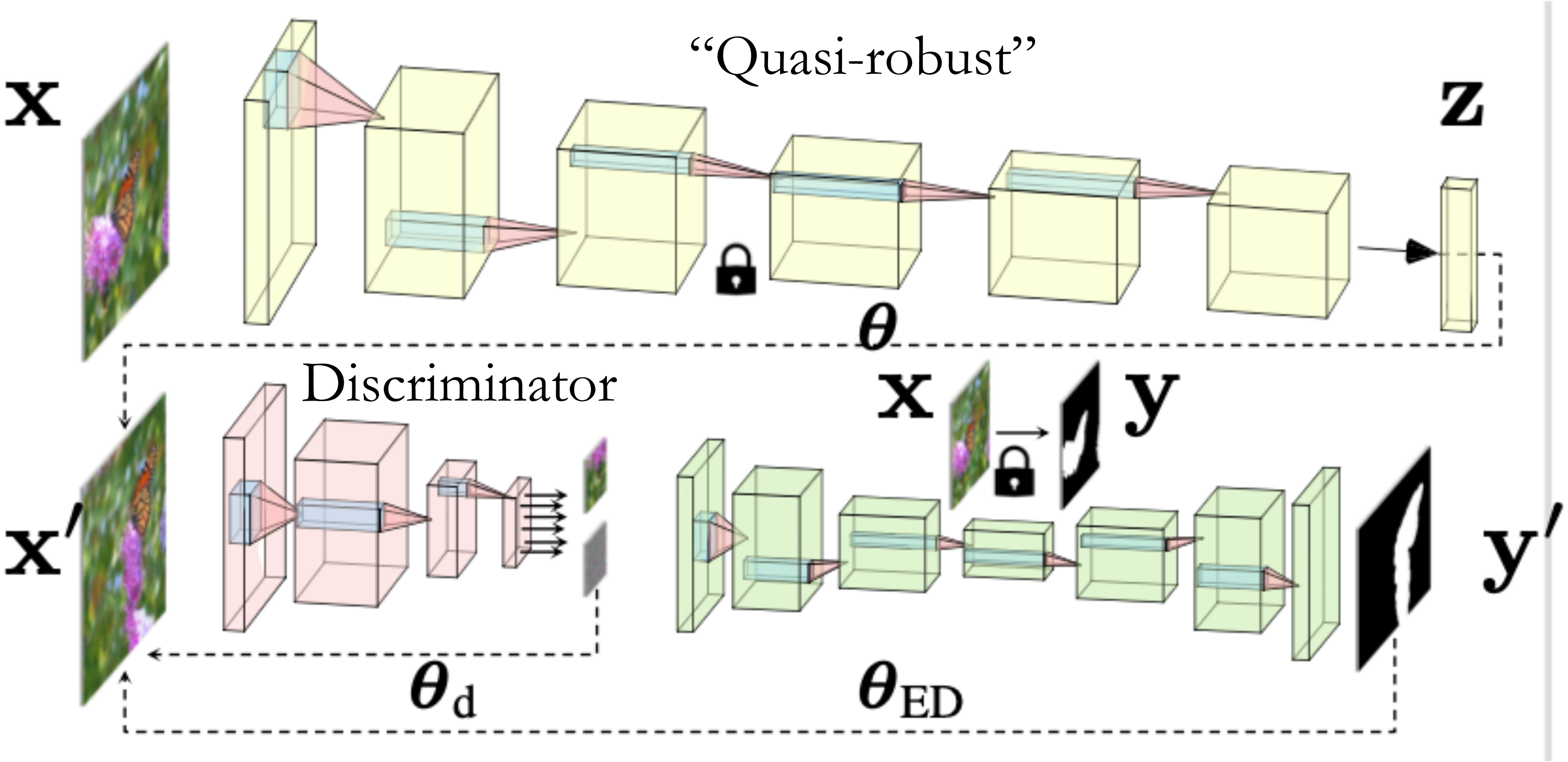
Method

$$
\begin{tikzcd}
\mathbf{x} \arrow[r, dashed, "\text{segment}"] \arrow[d, "\text{encode}"] \arrow[dd, bend right, dash] & \mathbf{y} \arrow[dd, dashed, "\text{edit}"] \\
\mathbf{z} \arrow[d, "\text{invert}"] & \\
\mathbf{x}' & \mathbf{y}'
\end{tikzcd}
$$

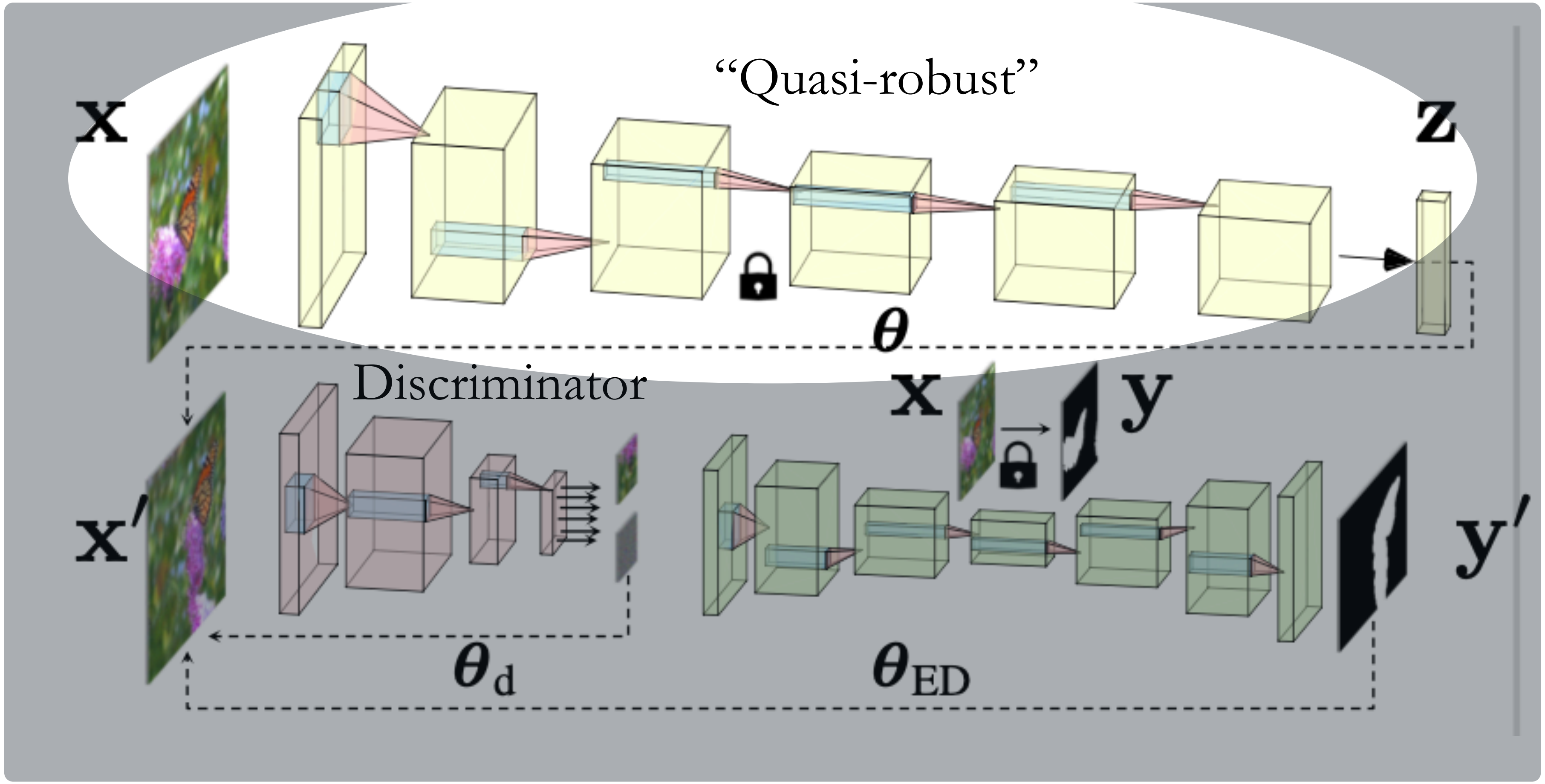
Method

$$
\begin{tikzcd}
\mathbf{x} \arrow[r, dashed, "\text{segment}"] \arrow[d, "\text{encode}"] \arrow[dd, bend right, no head] & \mathbf{y} \arrow[dd, dashed, "\text{edit}"] \\
\mathbf{z} \arrow[d, "\text{invert}"] & \\
\mathbf{x}' & \mathbf{y}' \arrow[l]
\end{tikzcd}
$$

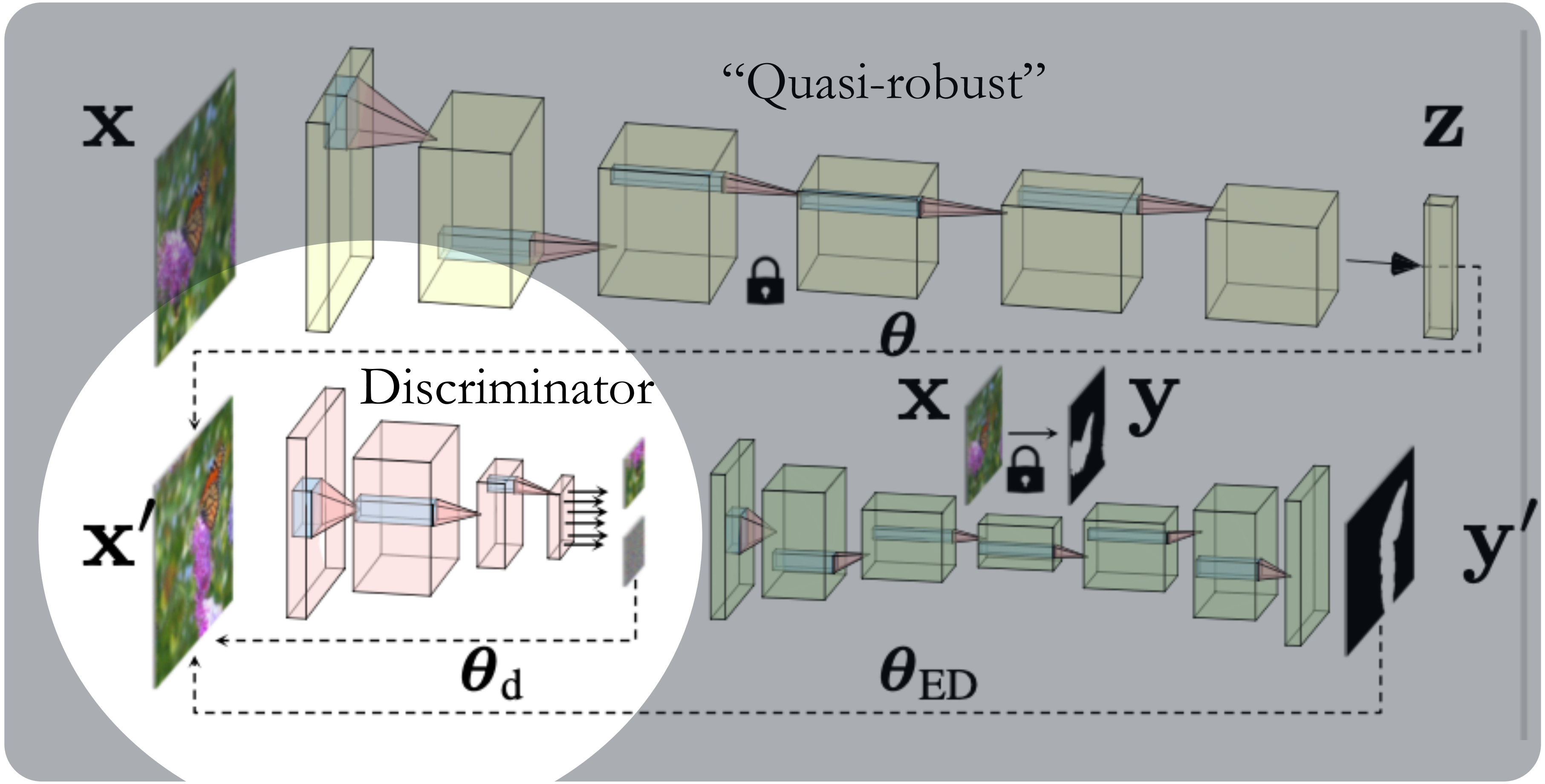
Manipulation Control



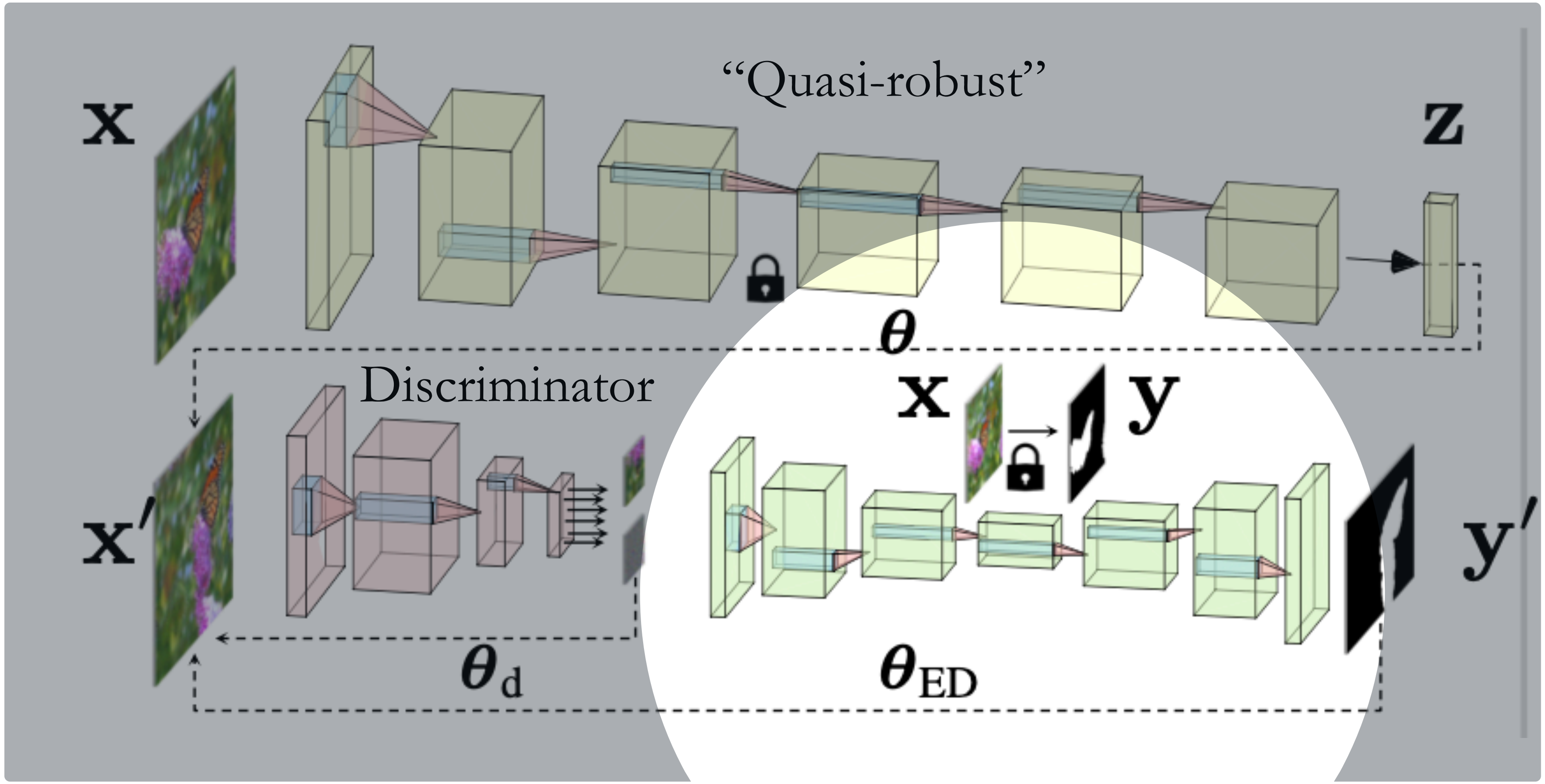
Manipulation Control



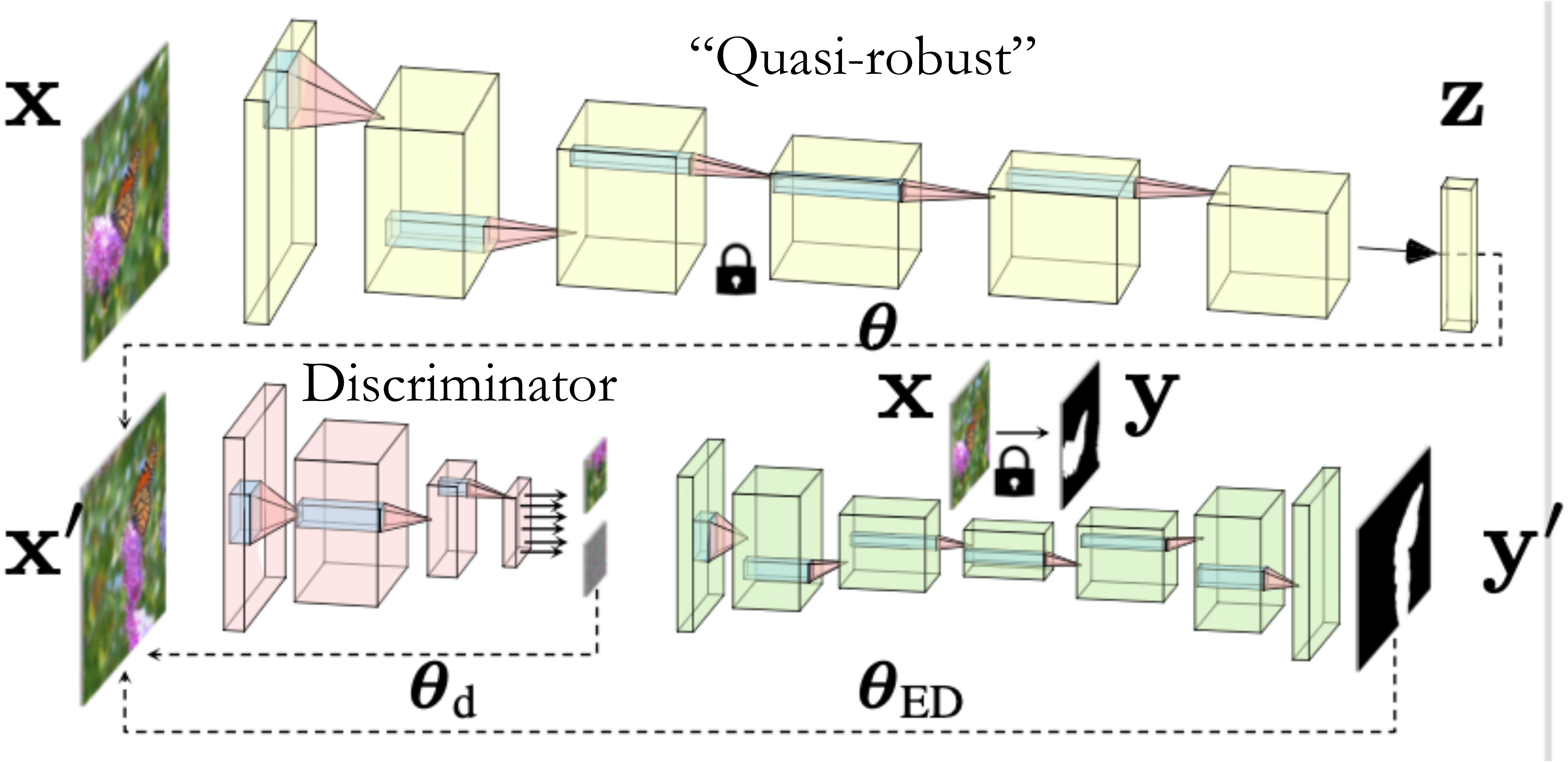
Manipulation Control



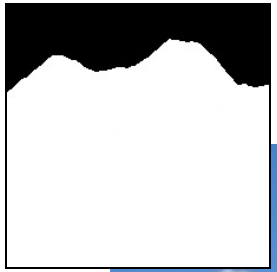
Manipulation Control



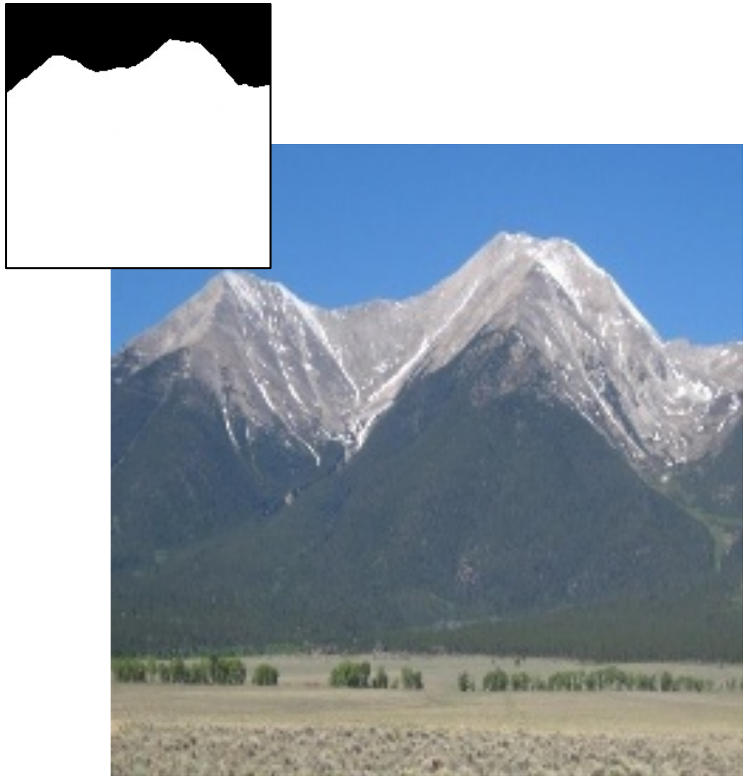
Manipulation Control



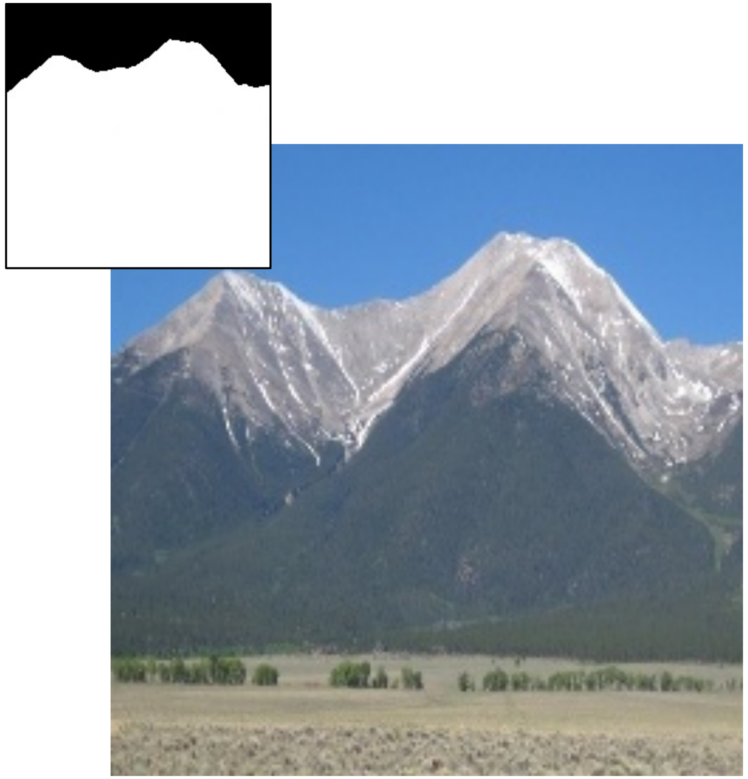
Manipulation Control – Non-Rigid Scene Deformation



Manipulation Control – Non-Rigid Scene Deformation

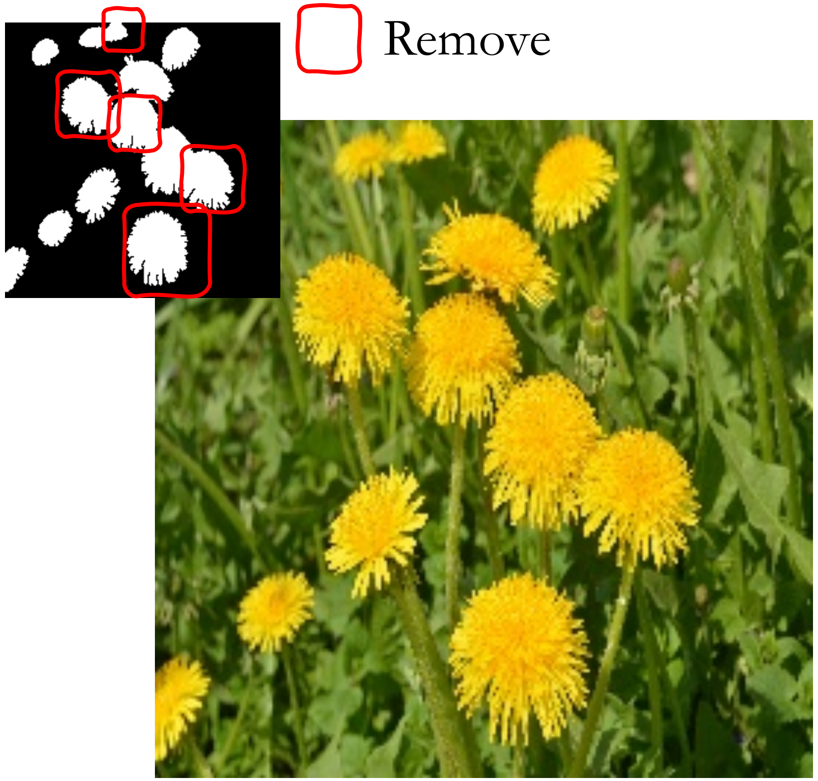


Manipulation Control – Non-Rigid Scene Deformation

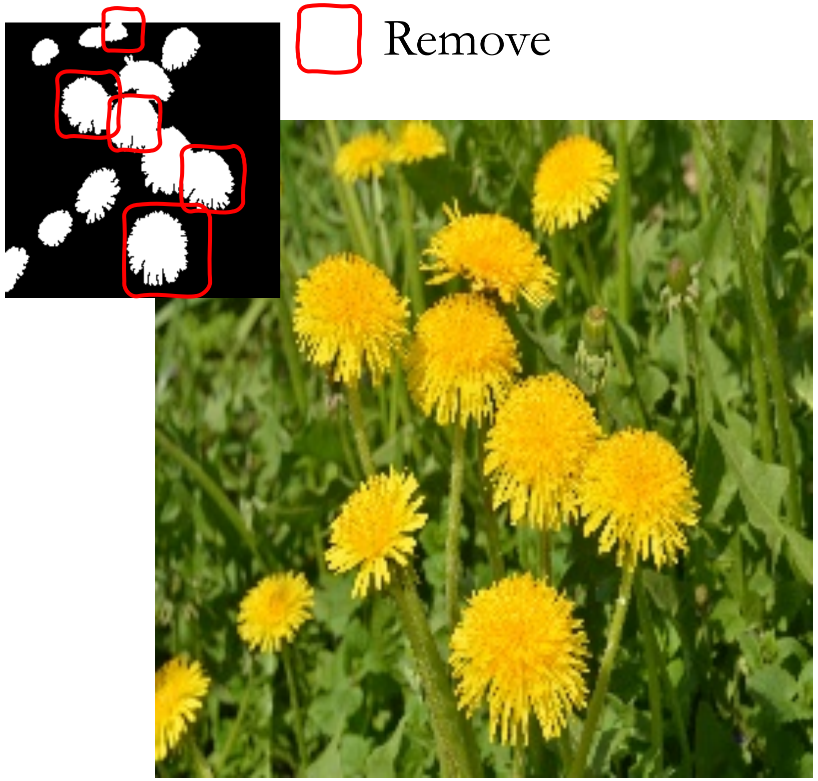


Manipulation Control – Copy/Move

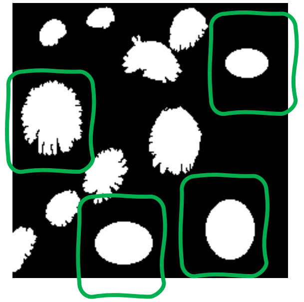
Manipulation Control – Copy/Move



Manipulation Control – Copy/Move



Remove



Add

Manipulation Control – Copy/Move

Qualitative Comparison

Input

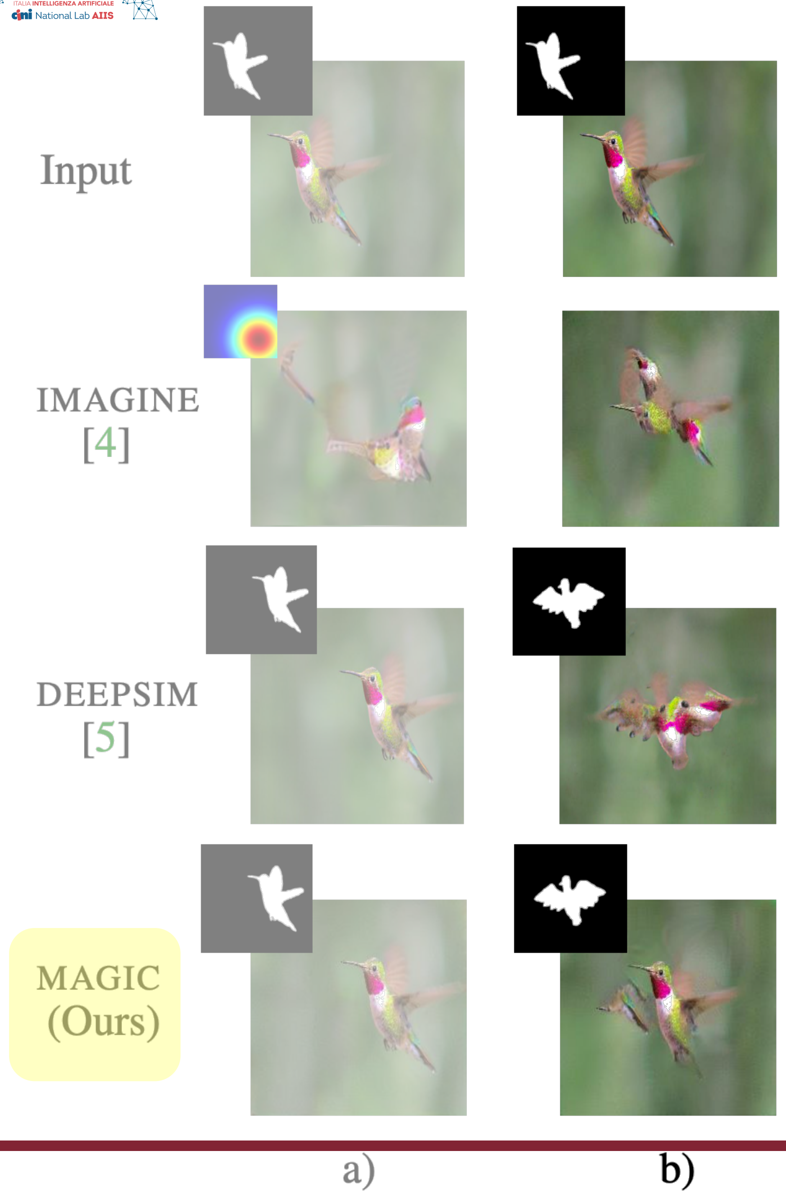
IMAGINE [4]

DEEPSIM [5]

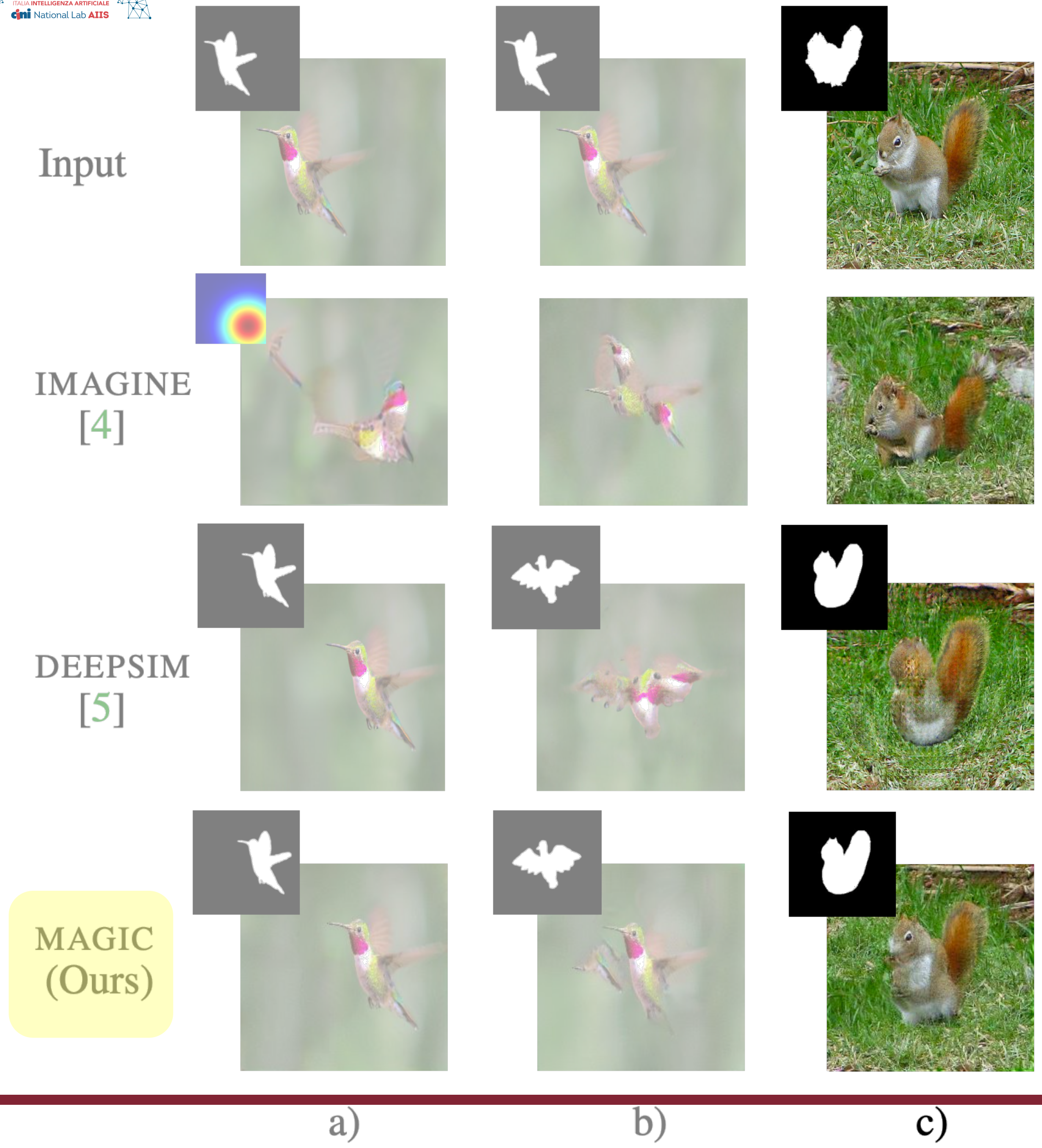
MAGIC (Ours)

a)

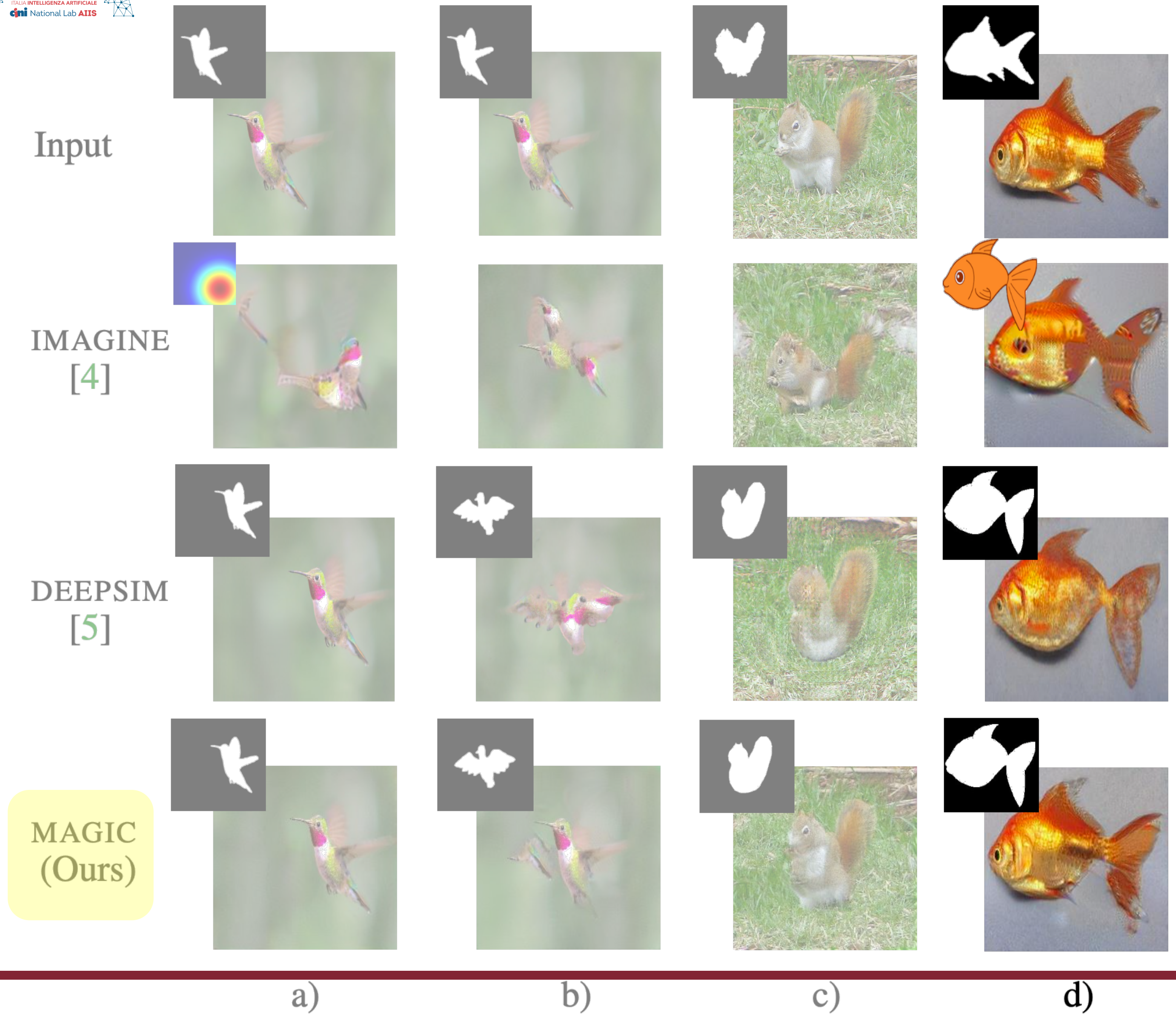
Qualitative Comparison



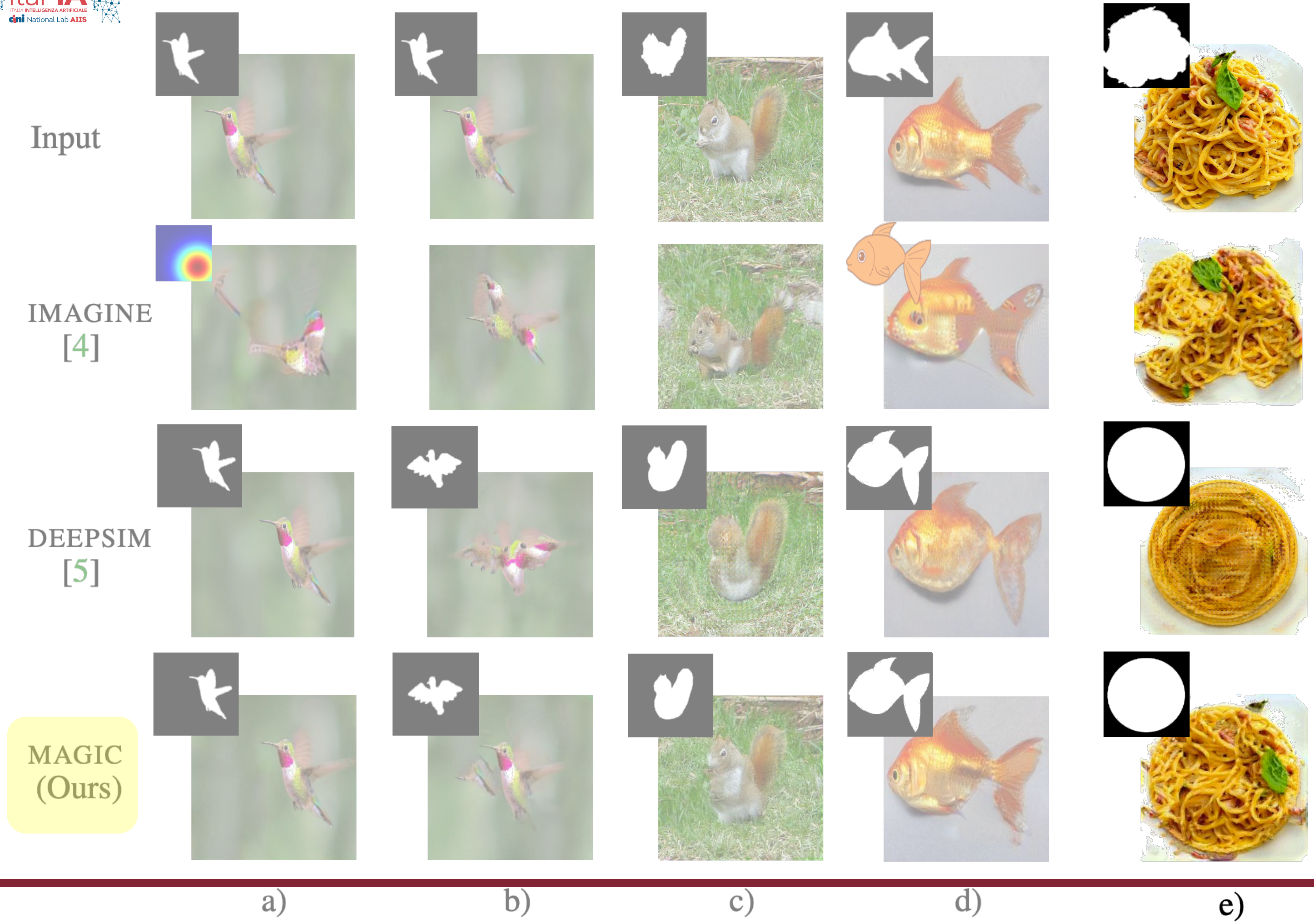
Qualitative Comparison



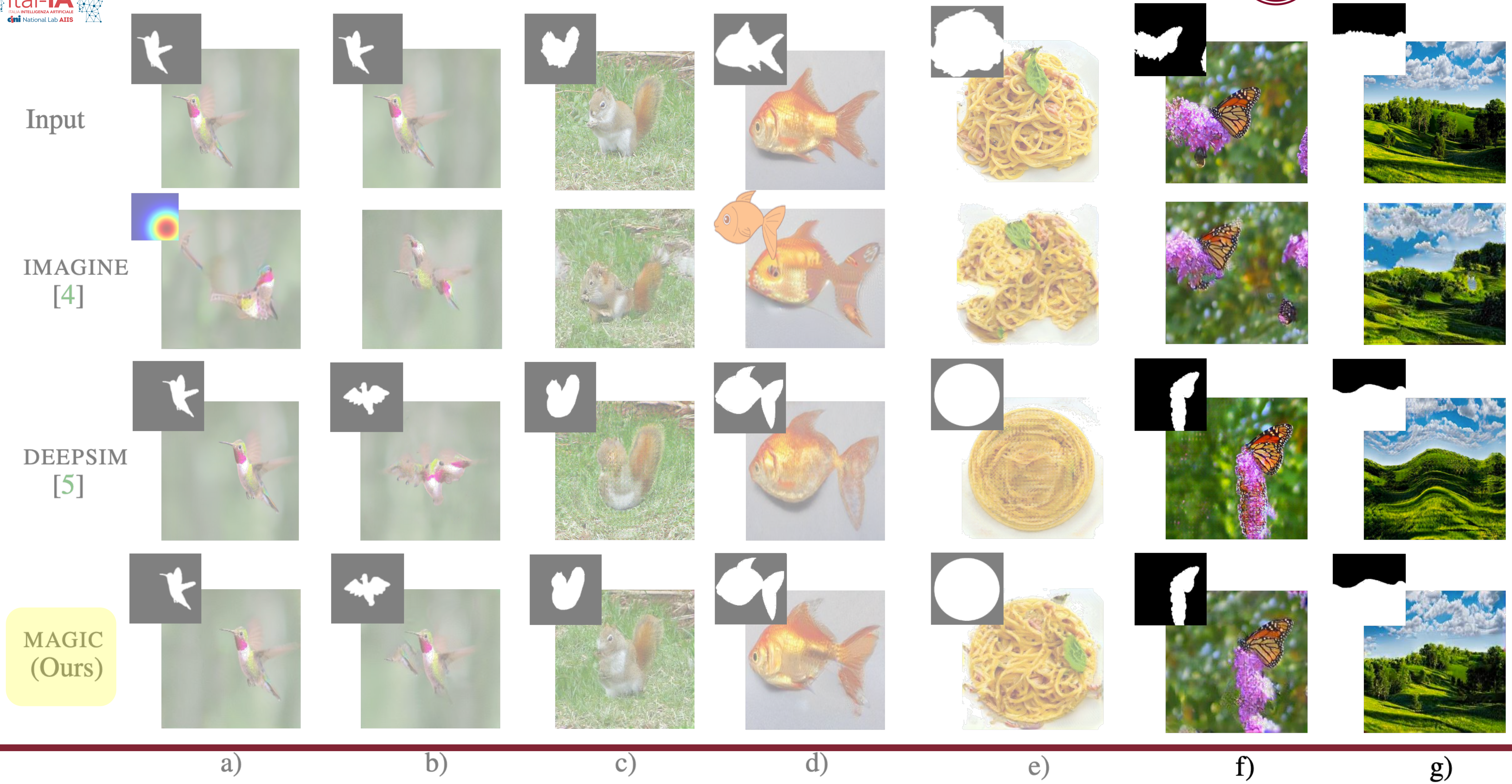
Qualitative Comparison



Qualitative Comparison



Qualitative Comparison



Future Work

Future Work

Better analyze High-Energy PGD

Investigate same but for targeted attacks

Future Work

Better analyze
High-Energy PGD

Investigate same but
for targeted attacks

Investigate Hybrid
Generative-Discriminative
Models

Future Work

Thank you!