ME-Net: Towards Effective Adversarial Robustness with Matrix Estimation

Yuzhe Yang

Guo Zhang, Dina Katabi, Zhi Xu

Poster #63

Massachusetts Institute of Technology

New defense: **ME-Net**

emphasizes *global structures* in images

Adversarial Examples

“pig”

Adversarial Examples

“pig”

Adversarial noise

Adversarial Examples

“pig”

Adversarial noise

“airliner”

Adversarial Examples

"pig"

Adversarial noise

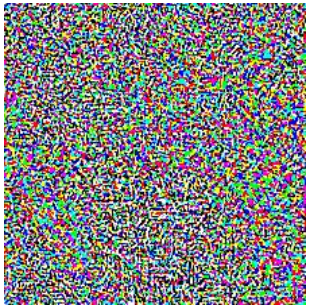
"airliner"

Highly structured

Idea: Destroy the Structure of Adversarial Noise

"pig"

Adversarial noise



Highly structured

"airliner"

Idea: Destroy the Structure of Adversarial Noise

"airliner"

Subsample

Idea: Destroy the Structure of Adversarial Noise

"airliner"

Subsample

But, subsampling destroys the structure of both adversarial noise and image

Idea: Destroy the structure of adversarial noise, but emphasize global structure of image!

"airliner"

Subsample

Idea: Destroy the structure of adversarial noise, but emphasize global structure of image!

"airliner"

Subsample

Images are known to be low rank!

Idea: Destroy the structure of adversarial noise, but emphasize global structure of image!

"airliner"

Subsample

Images are known to be low rank!

Matrix Estimation can be used to recover global structures in an image

Idea: Destroy the structure of adversarial noise, but emphasize global structure of image!

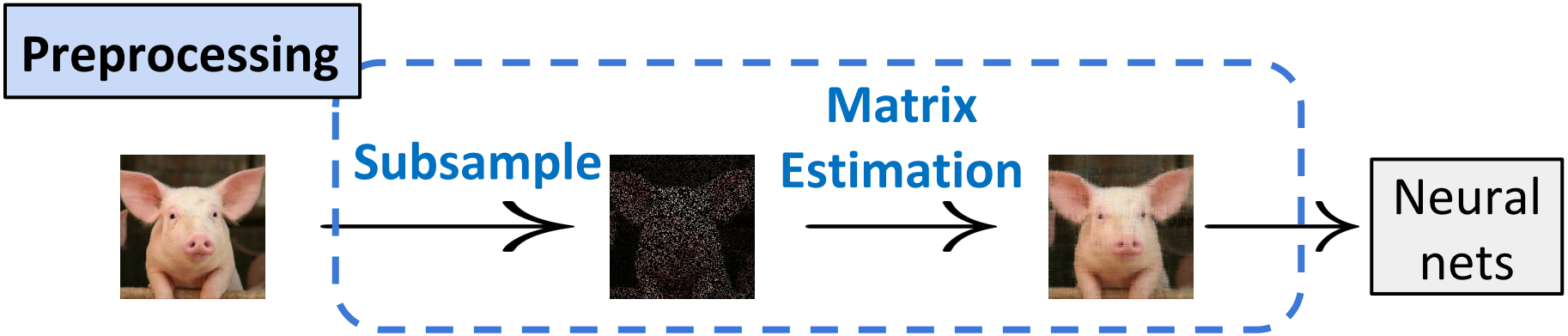
"airliner"

Subsample

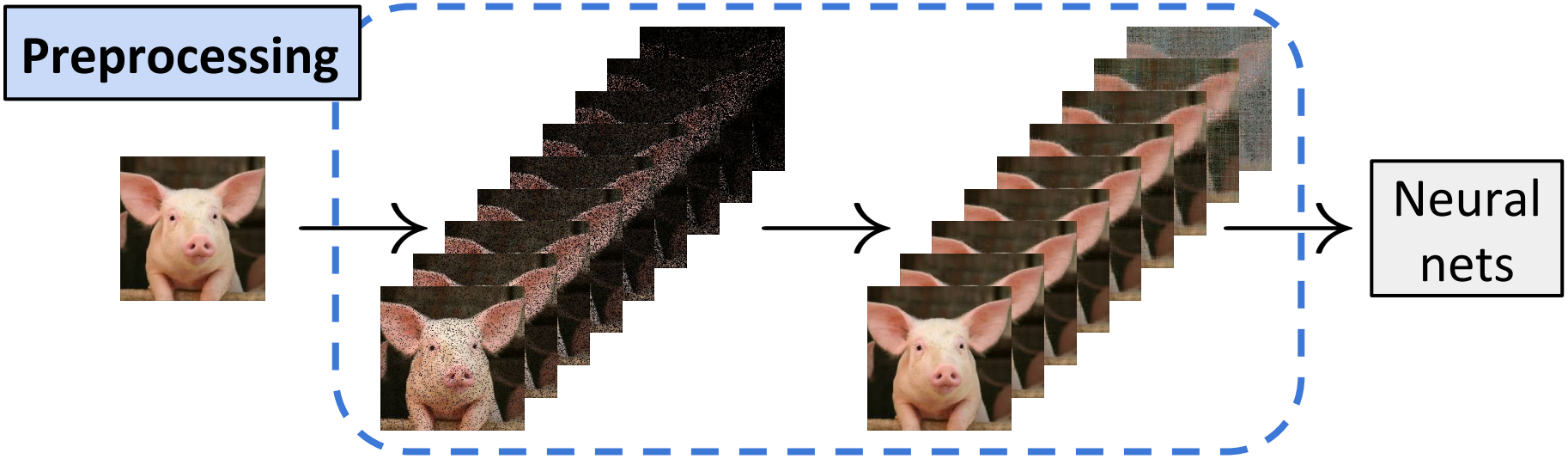
Matrix Estimation

"pig"

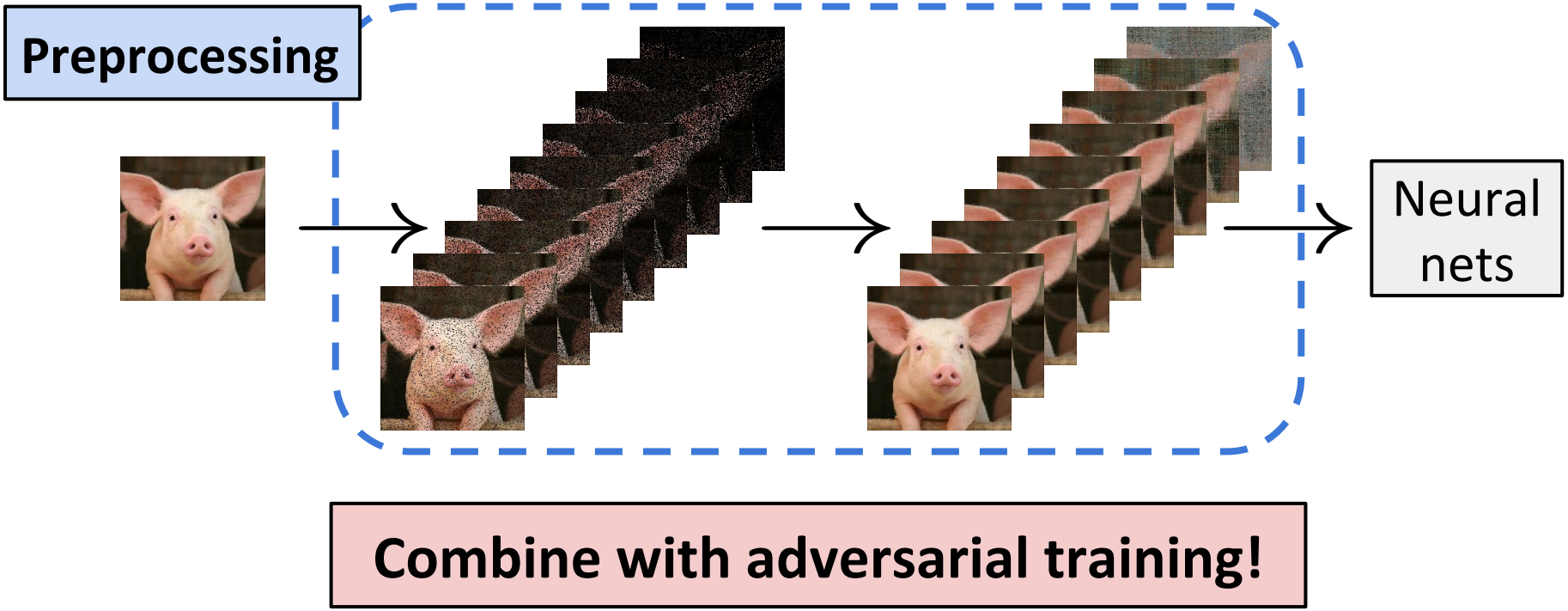
ME-Net



ME-Net



ME-Net


Preprocessing
Neural nets
Combine with adversarial training!

CIFAR-10 Black-Box Attacks

transfer-based | decision-based | score-based

	CW	FGSM	PGD	Boundary	SPSA
Vanilla					
Madry et al.					
ME-Net					

CIFAR-10 Black-Box Attacks

	transfer-based			decision-based	score-based
	CW	FGSM	PGD	Boundary	SPSA
Vanilla	8.9%	24.8%	7.6%	3.5%	1.4%
Madry et al.	78.7%	67%	64.2%	61.9%	47.0%
ME-Net	**93.6%**	**92.2%**	**91.8%**	**87.4%**	**93.0%**

High robustness against black-box attacks!

White-box Attacks

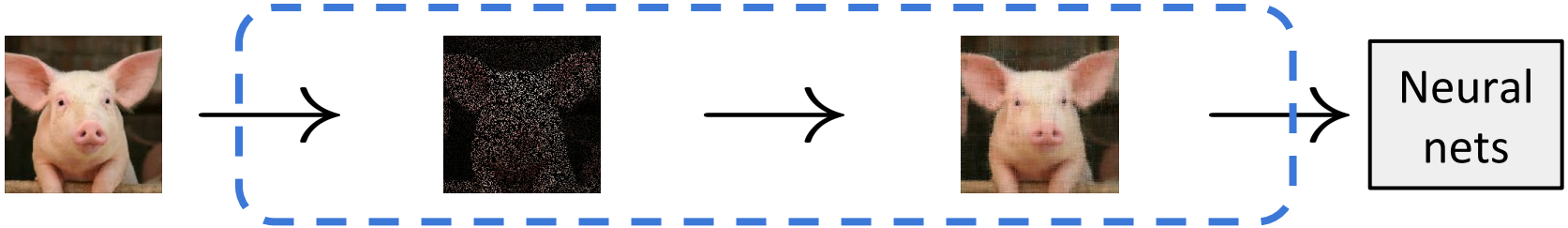
	MNIST	CIFAR-10	SVHN	Tiny-ImageNet (Top-1)
Madry et al.				
ME-Net + SGD				
ME-Net + Adv.				

White-box Attacks

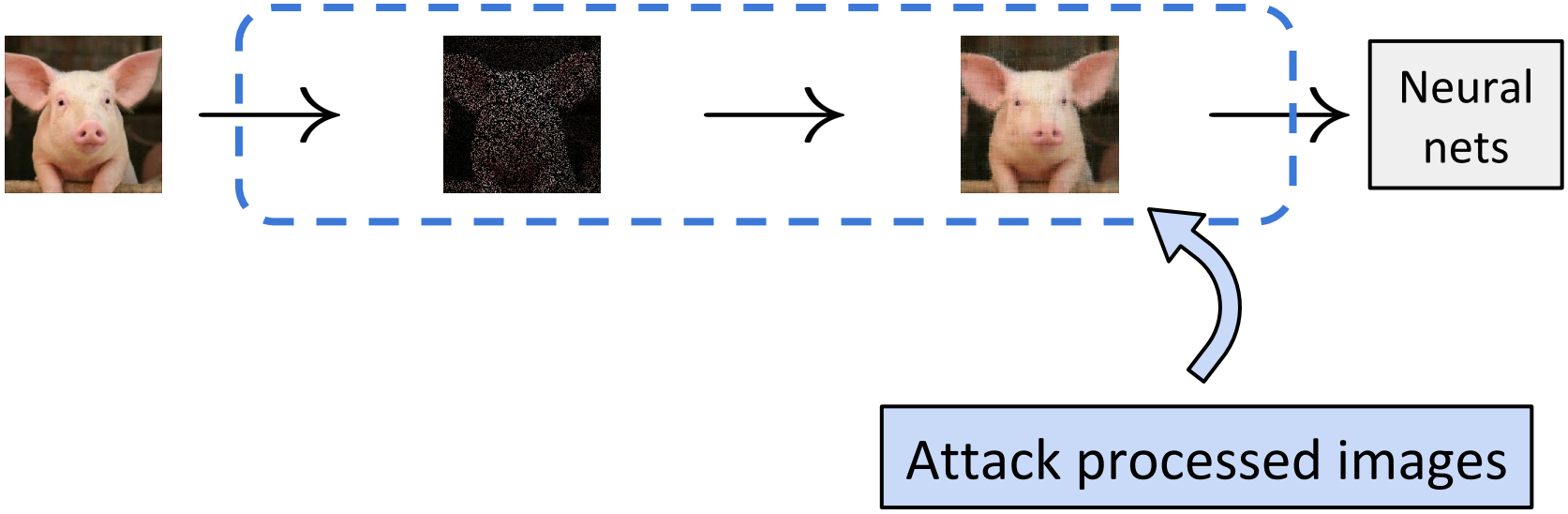
	MNIST	CIFAR-10	SVHN	Tiny-ImageNet (Top-1)
Madry et al.	**91.6%**	45.0%	47.1%	22.1%
ME-Net + SGD	82.6%	40.8%	43.4%	18.9%
ME-Net + Adv.	91.0%	**52.8%**	**69.4%**	**28.5%**

Improve white-box robustness when combined with AT!

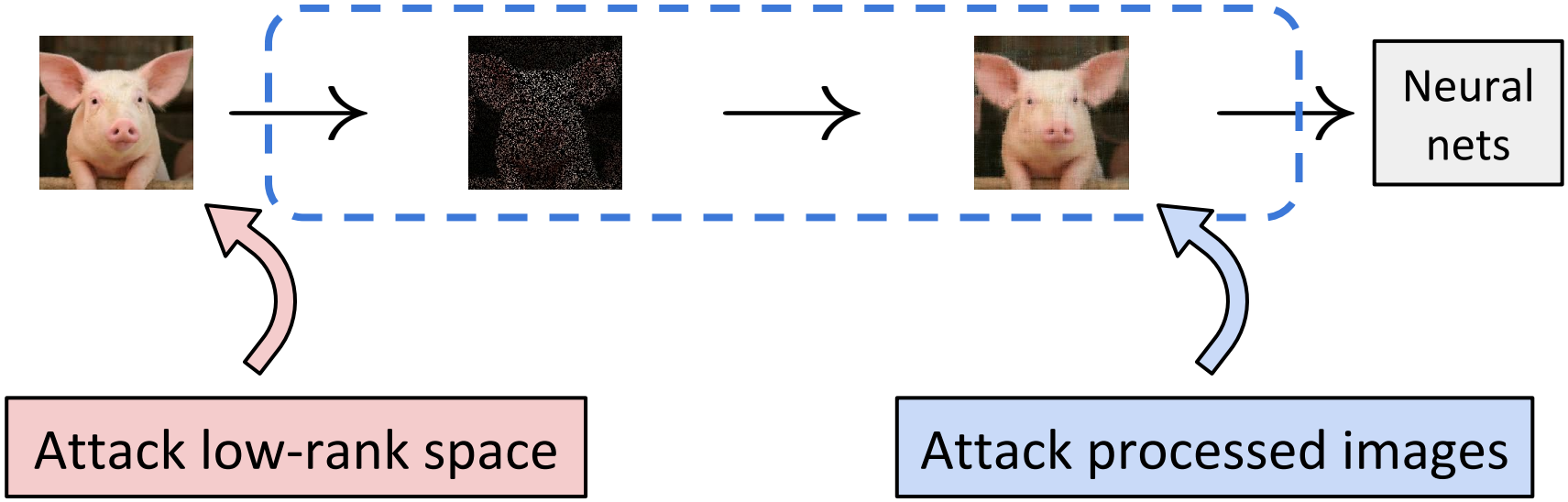
Customized Adaptive Attacks



Also Robust to Customized Adaptive Attacks [Results in Paper]



Also Robust to Customized Adaptive Attacks [Results in Paper]



Please visit our poster #63

Source Code

Project Page