The Best Defense is a Good Offense: Adversarial Augmentation against Adversarial Attacks (A^5)

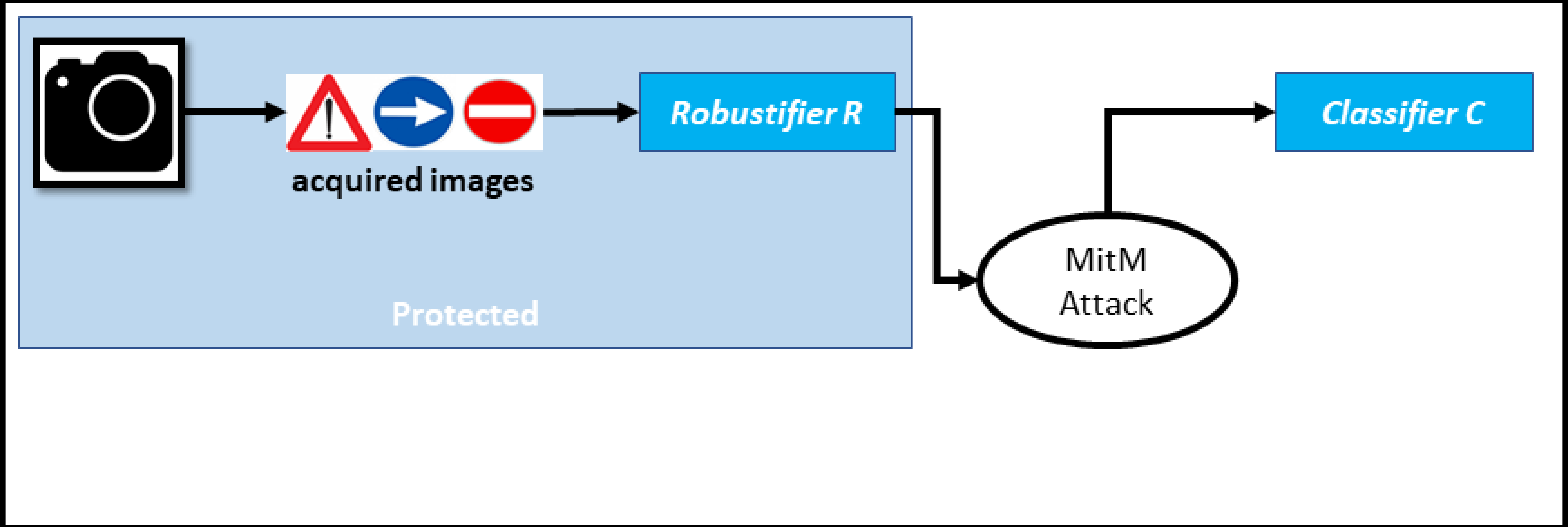
Iuri Frosio, Jan Kautz
NVIDIA

CVPR 2023

Poster ID 388
TUE-AM-388

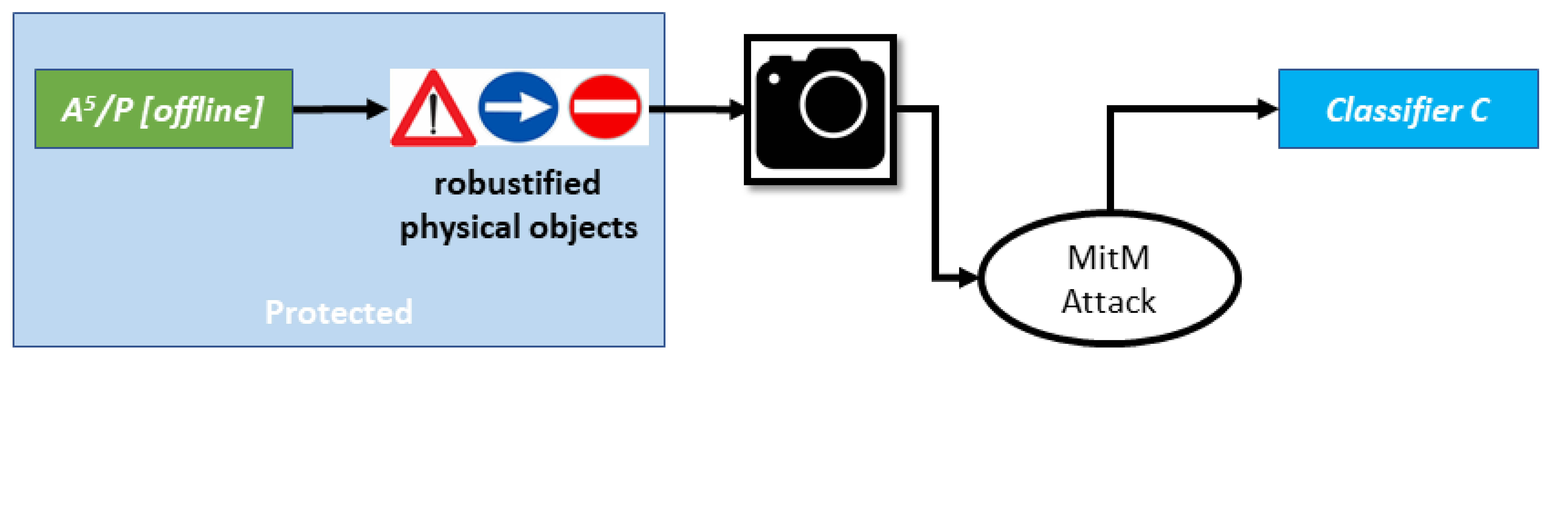
Adversarial Augmentation against Adversarial Attacks (A^5)

A full framework for PREEMPTIVE, CERTIFIED protection



Adversarial Augmentation against Adversarial Attacks (A^5)

A full framework for PREEMPTIVE, CERTIFIED protection

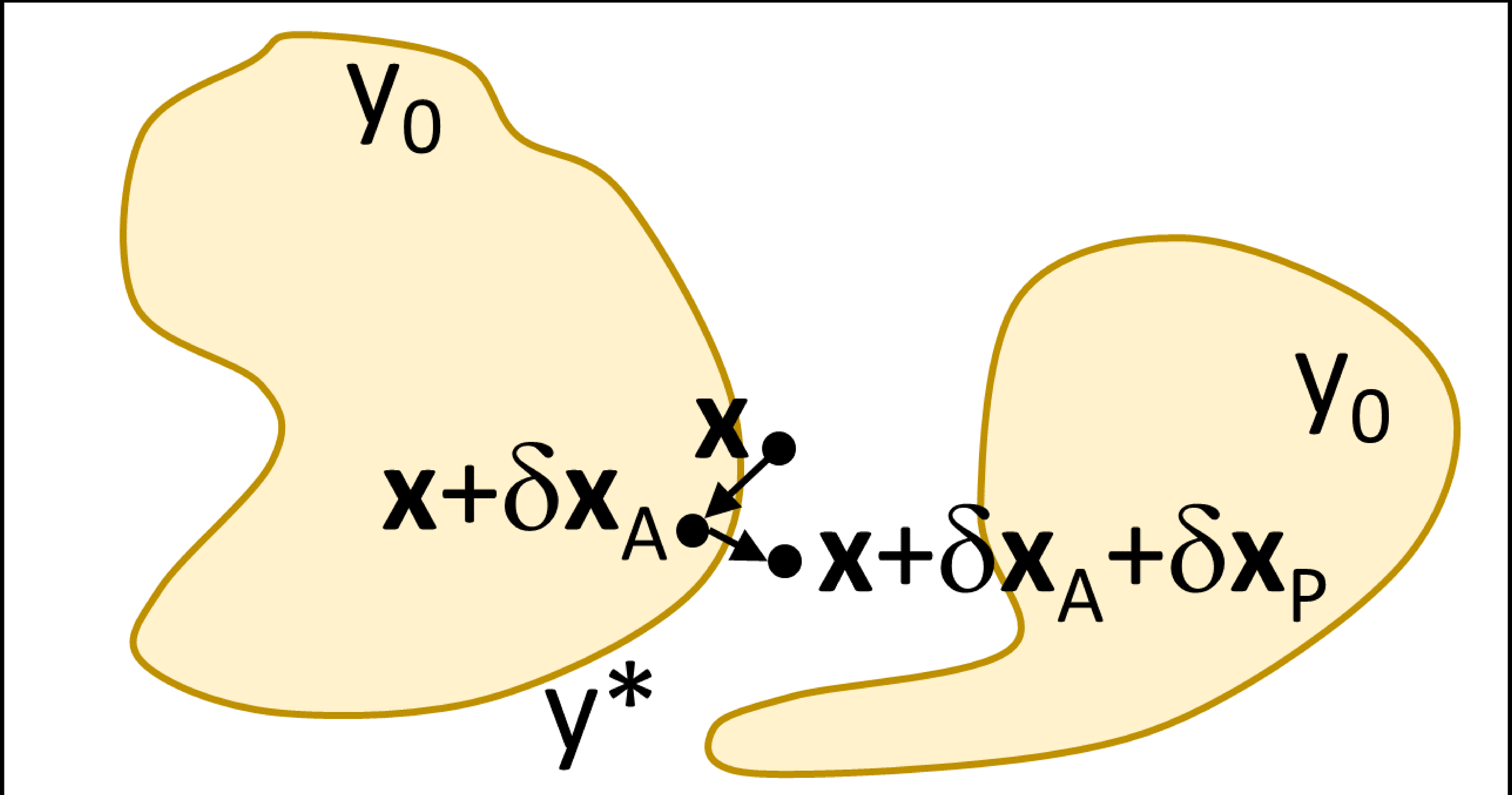


Adversarial Augmentation against Adversarial Attacks (A^5)

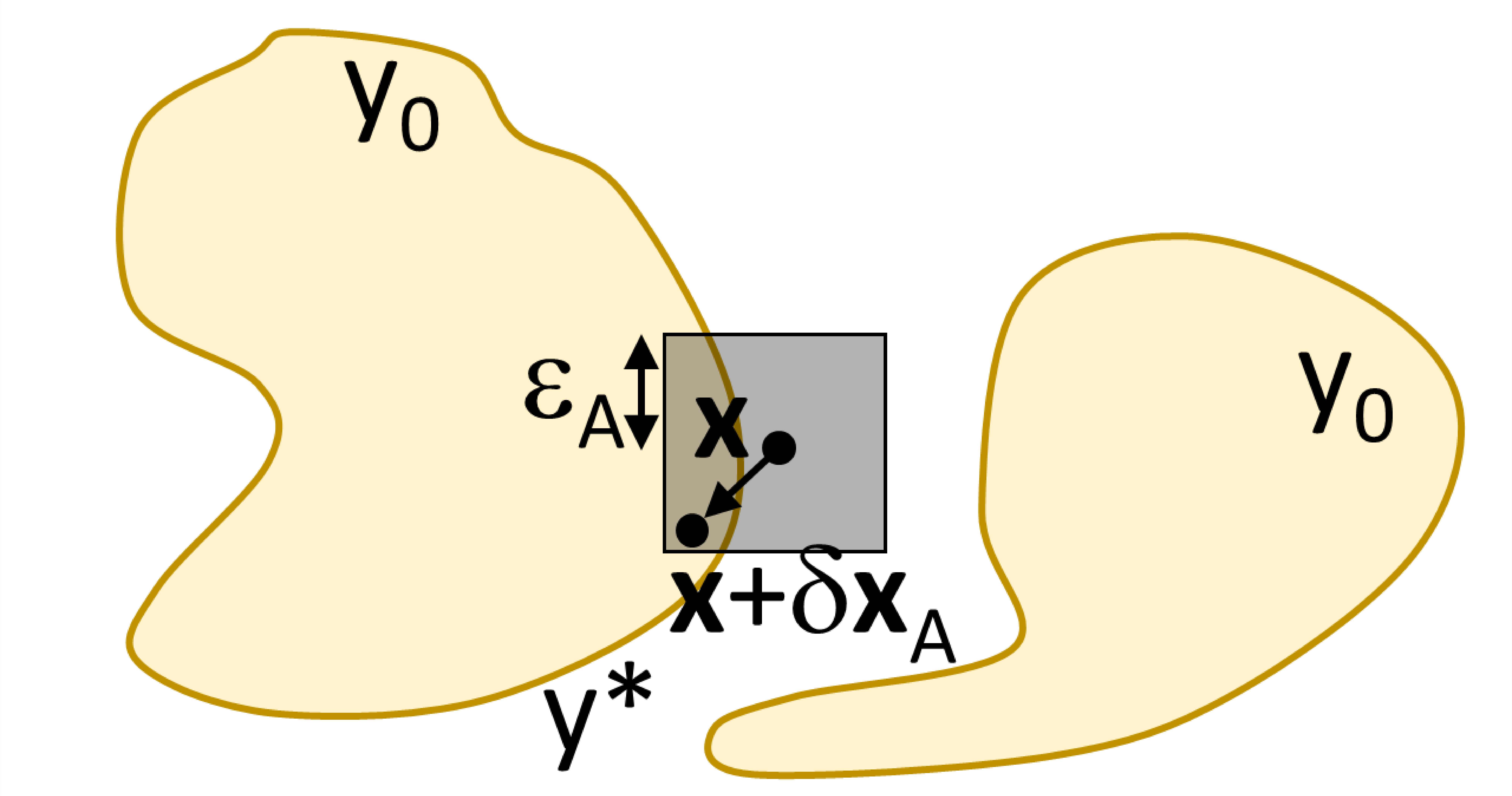
A full framework for PREEMPTIVE, CERTIFIED protection

w
$w+\delta w_D$
A^5/P, A^5/PC, A^5/O
Protected
$w+\delta w_D+\delta w_A$
Physical Attack *
Acquisition A
$x=A(w+\delta w_D+\delta w_A)$
Robustifier R
A^5/R, A^5/RC
$x+\delta x_D$
Protected
$x+\delta x_D+\delta x_A$
MitM Attack
Classifier C
A^5/RC, A^5/PC
Protected
data
...
algorithm
...
DNN

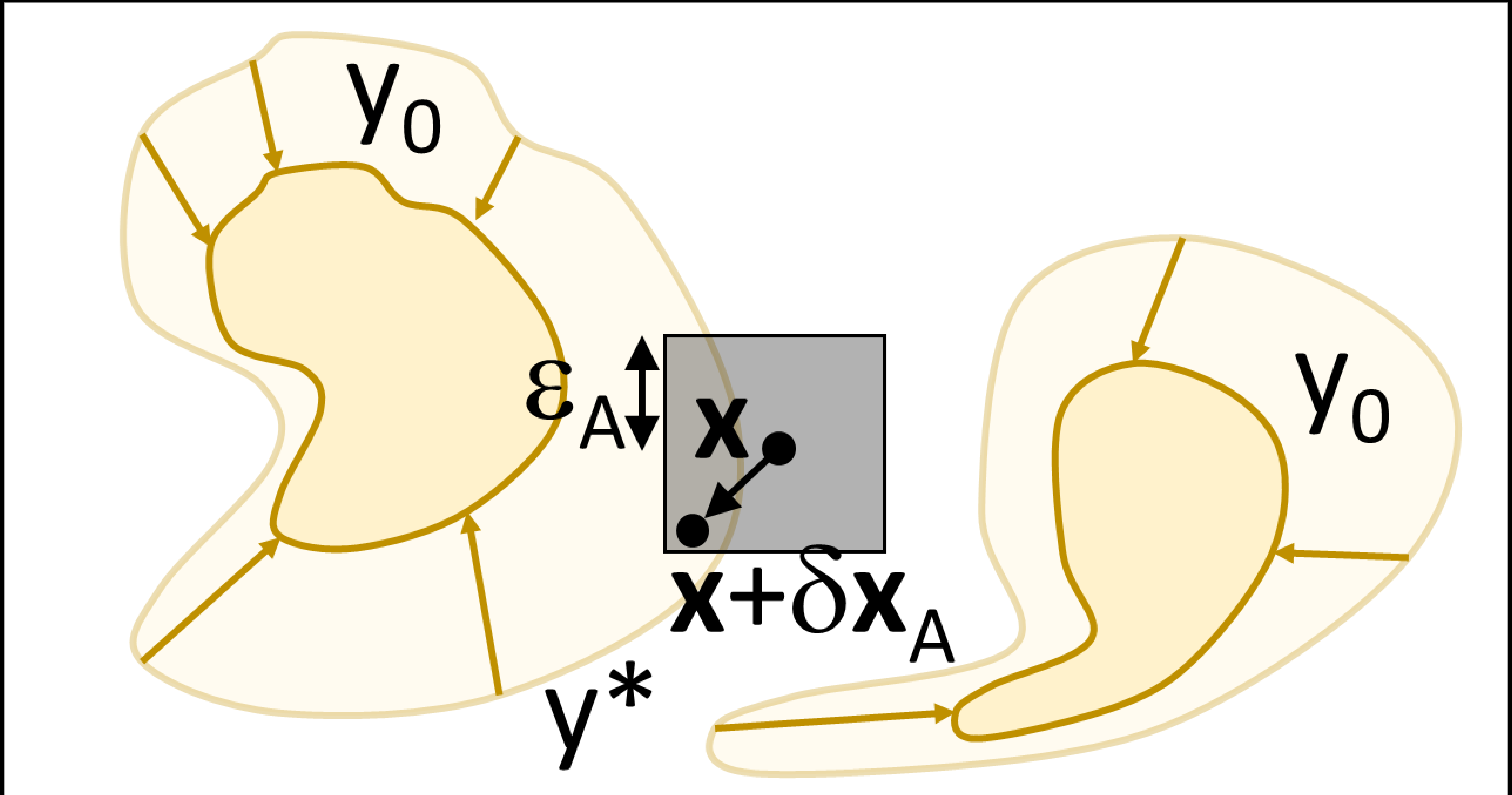
Other defense methods: purification (and randomization)


y_0
$\mathbf{x}$
$\mathbf{x}+\delta\mathbf{x}_A$
$\mathbf{x}+\delta\mathbf{x}_A+\delta\mathbf{x}_P$
y_0
y^*

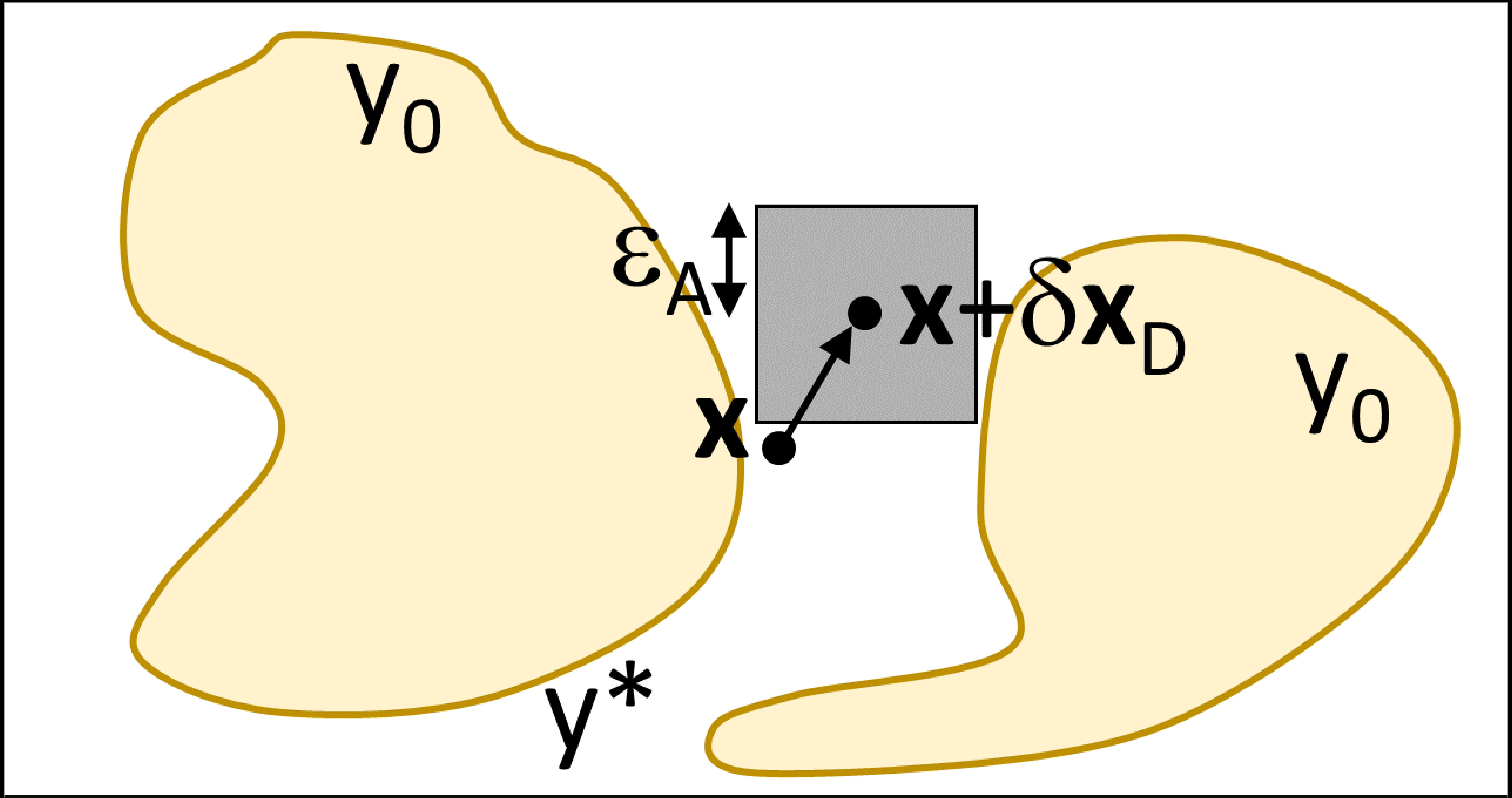
Bound computation



Adversarial and certified training



A^5



Configurations: Offline (A^5/O)

w

$w+\delta w_D$

A^5/O

Protected

$A() = normalize()$

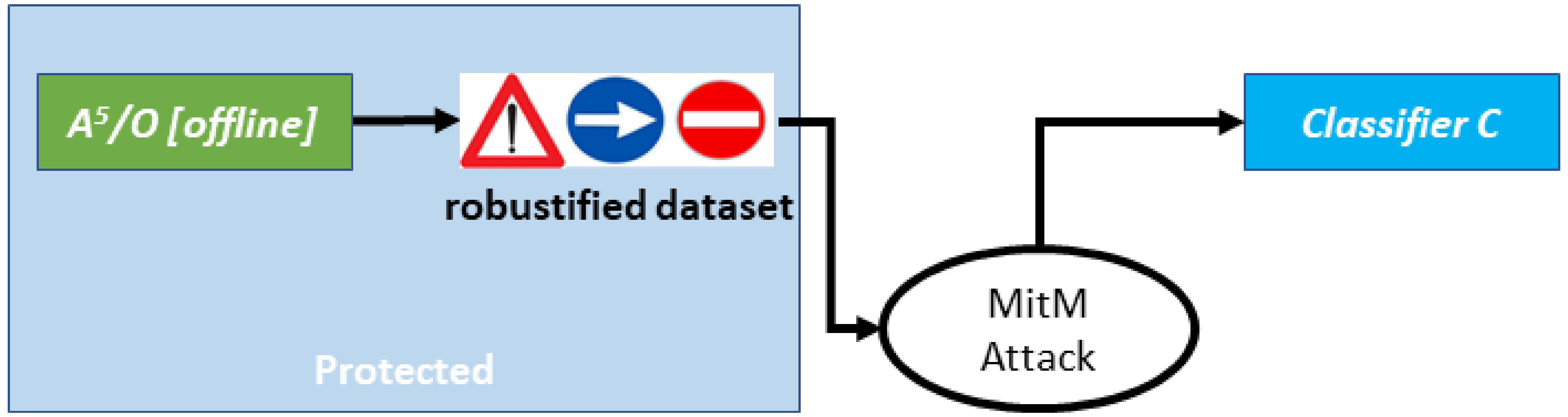
$x+\delta x_D=A(w+\delta w_D)$

Protected

$x+\delta x_D+\delta x_A$

MitM Attack

Classifier C



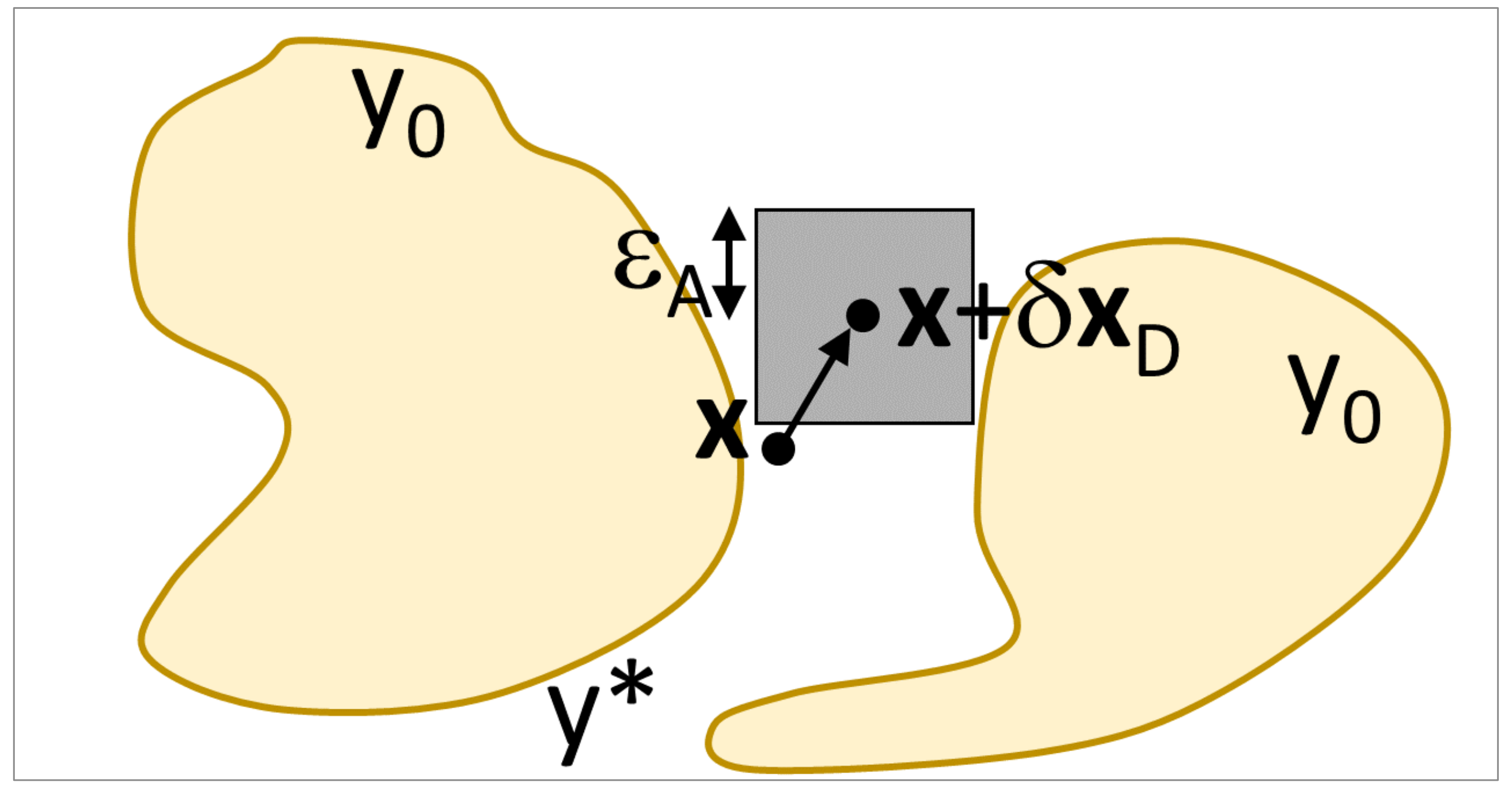
A^5/O [offline]

robustified dataset

Protected

MitM Attack

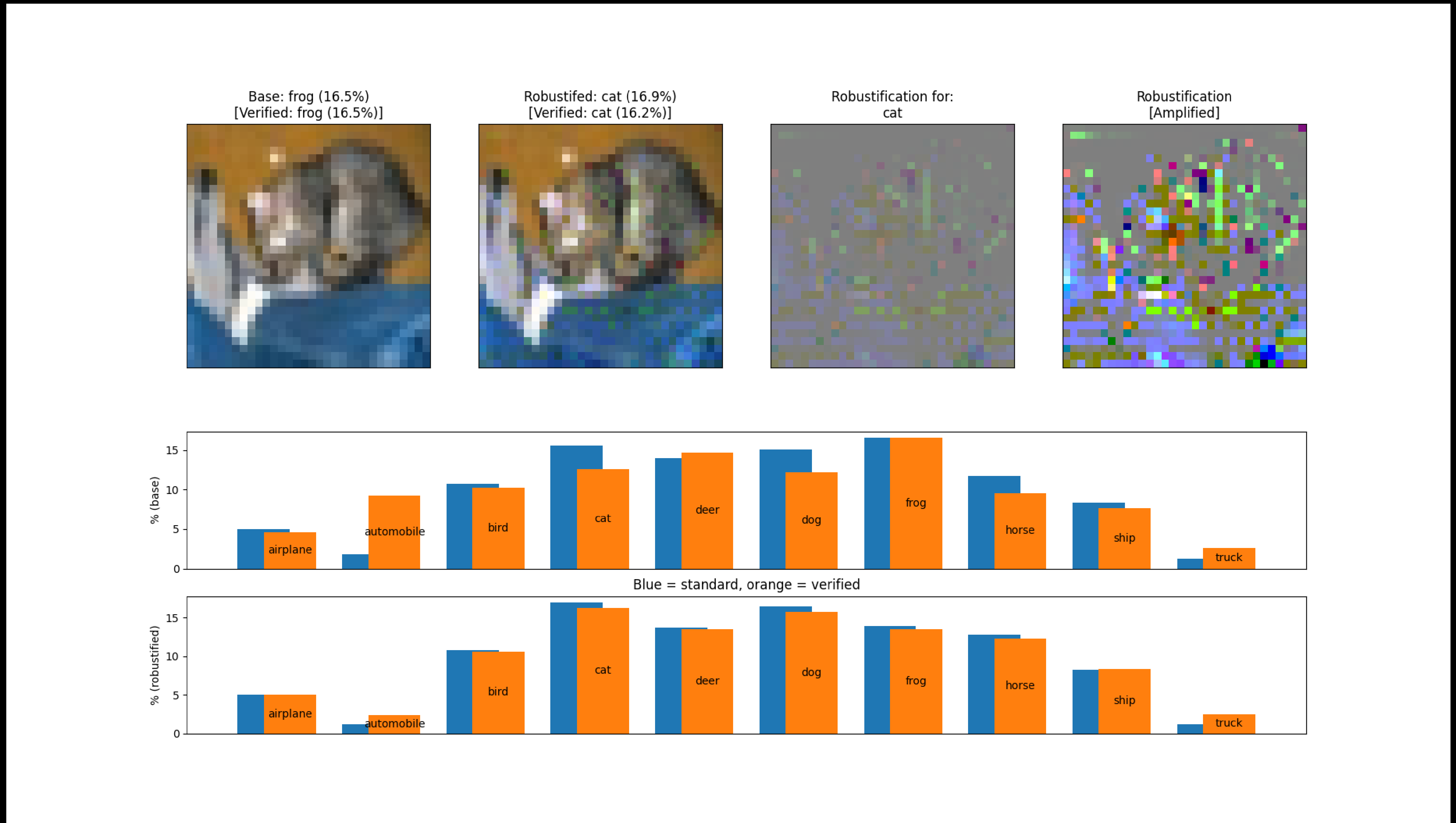
Classifier C



Configurations: Offline (A^5/O)

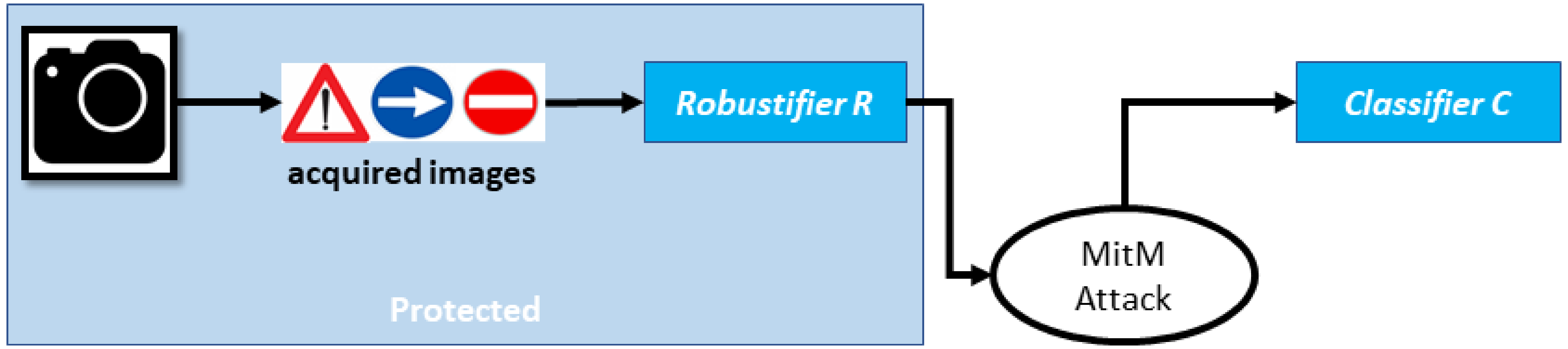
Results on CIFAR10, attack 8/255

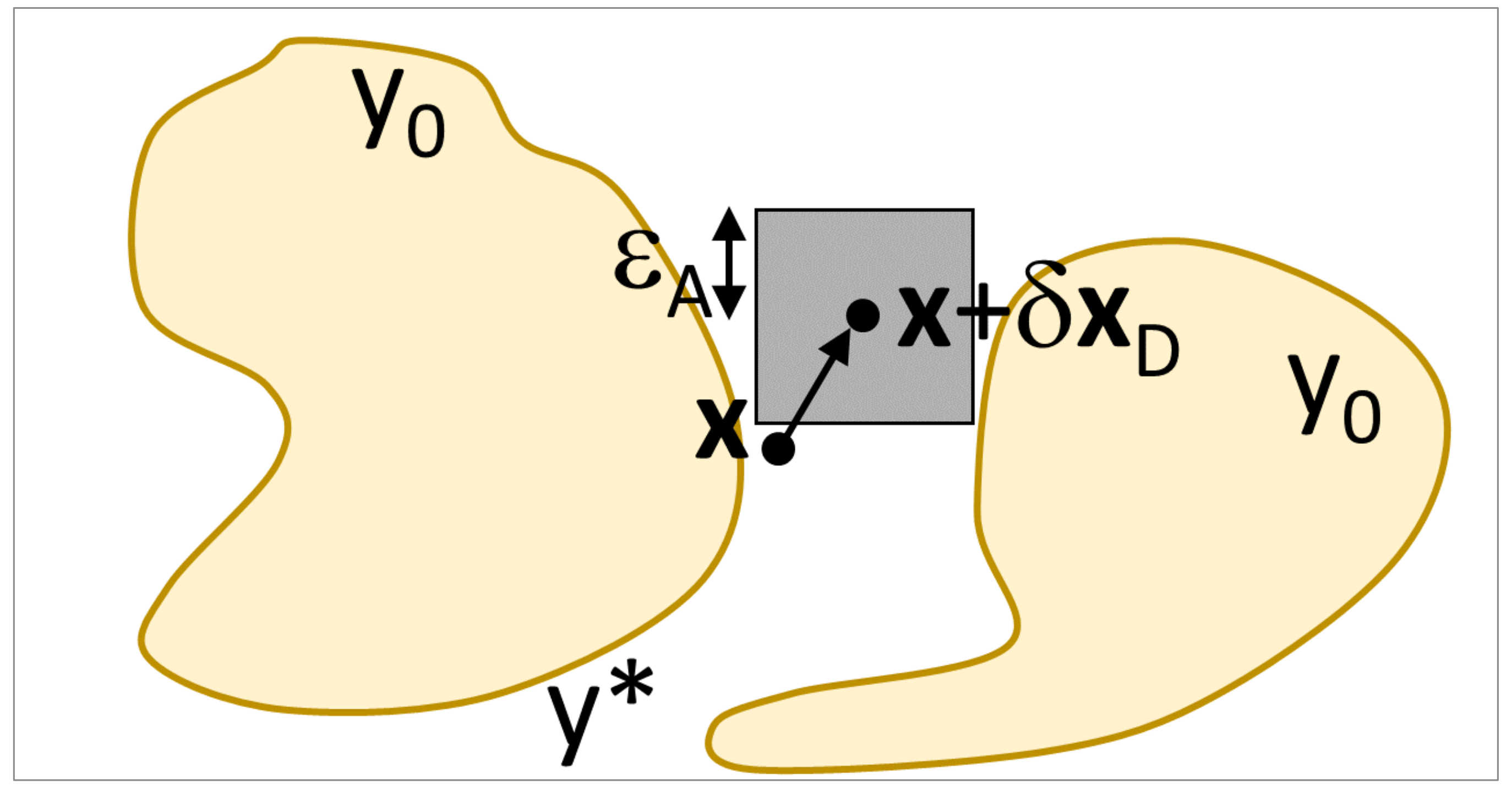
Method	Clean error	Certified error
CROWN-IBP	54.02%	66.94%
A^5/O	45.68%	49.74%



Configurations: Robustifier (A⁵/R)

w
Protected
A() = normalize()
x=A(w)
Robustifier R
A⁵/R
x+δx_D
Protected
x+δx_D+δx_A
MitM Attack
Classifier C

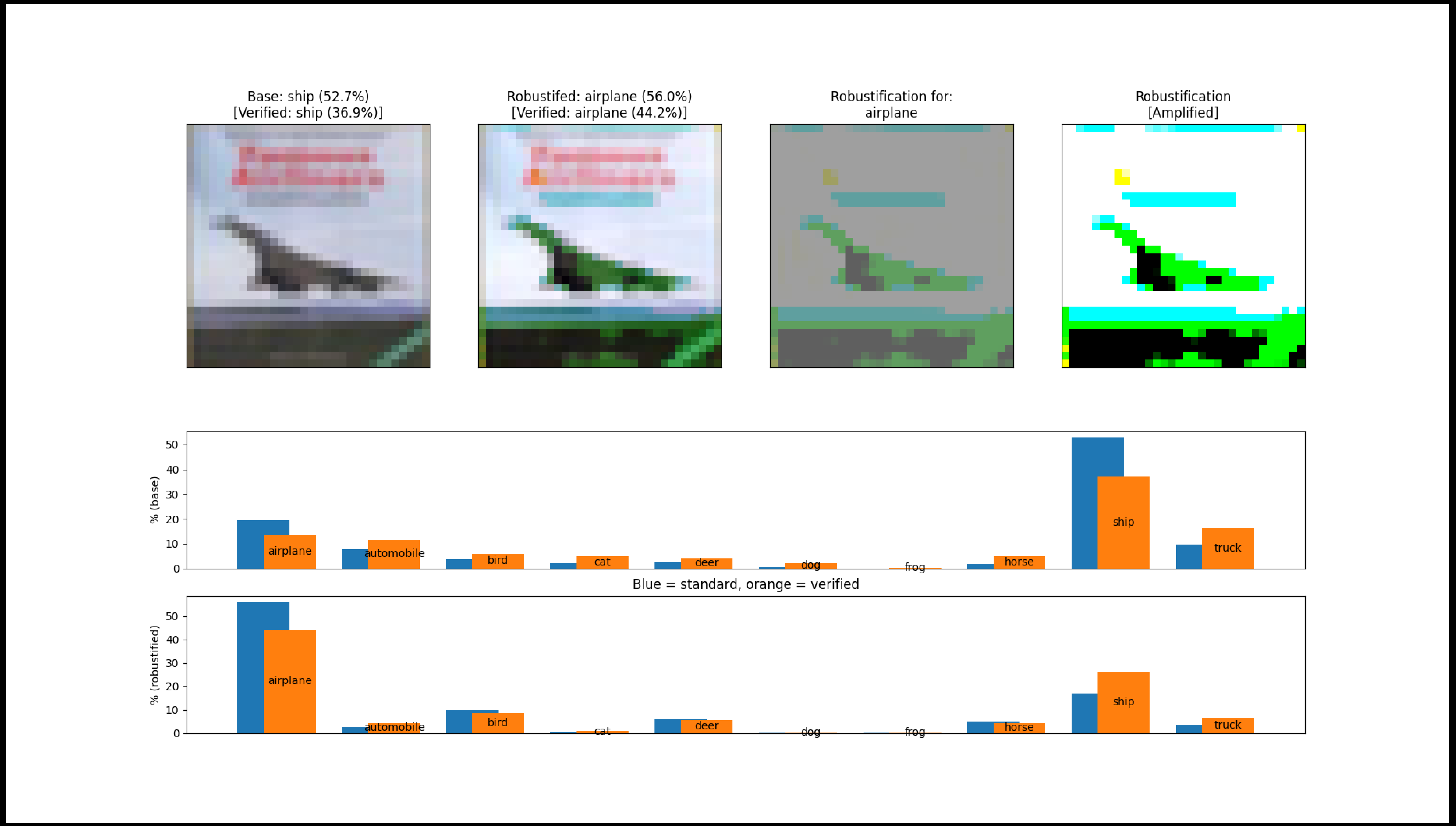




Configurations: Robustifier (A^5/R)

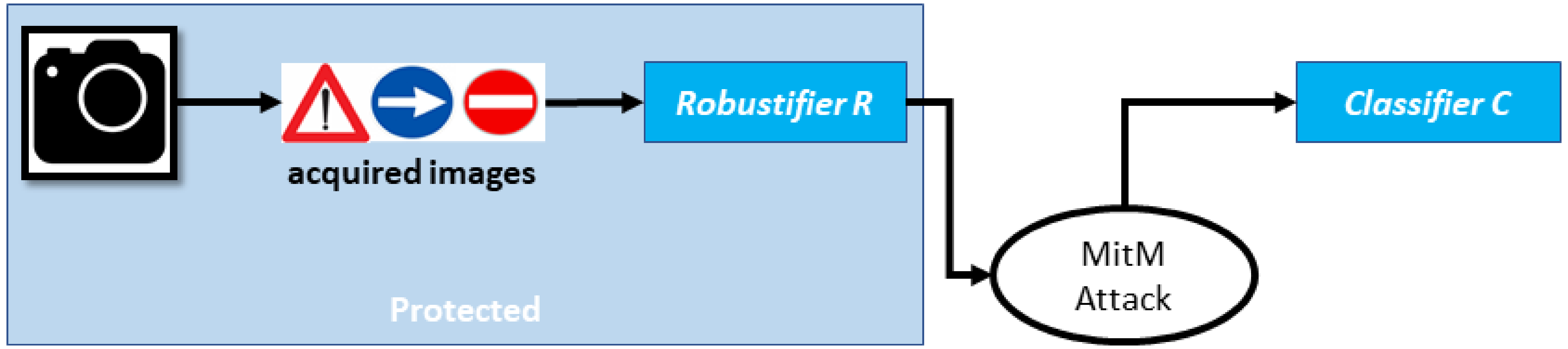
Results on CIFAR10, attack 8/255

Method	Clean error	Certified error
CROWN-IBP	54.02%	66.94%
A^5/O	45.68%	49.74%
A^5/R	50.91%	57.95%



Configurations: Robustifier and Classifier (A^5/RC)

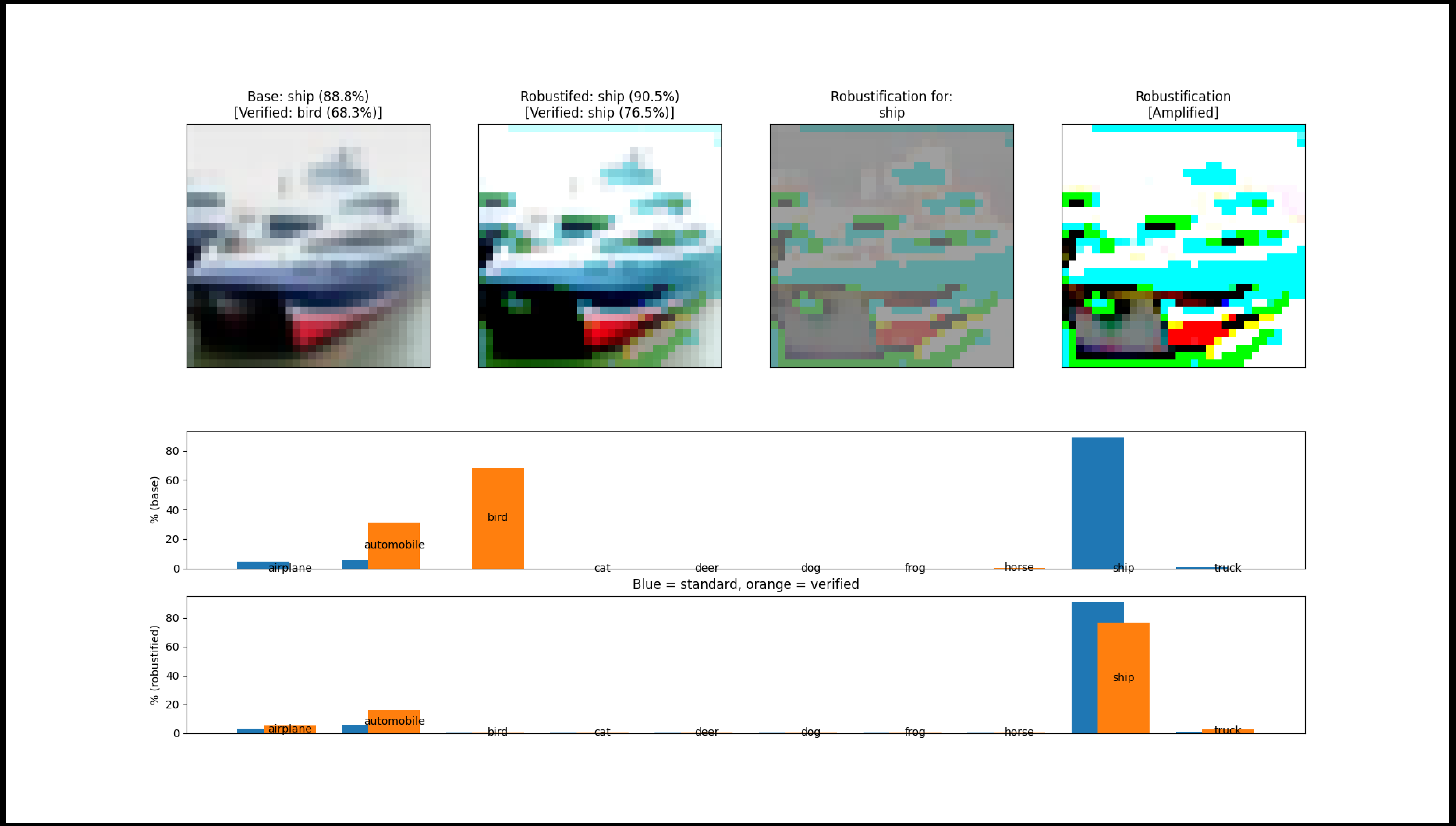
w
Protected
$A() = normalize()$
$x=A(w)$
Robustifier R
A^5/RC
$x+\delta x_D$
Protected
$x+\delta x_D+\delta x_A$
MitM
Attack
Classifier C
A^5/RC
Protected



Configurations: Robustifier and Classifier (A^5/RC)

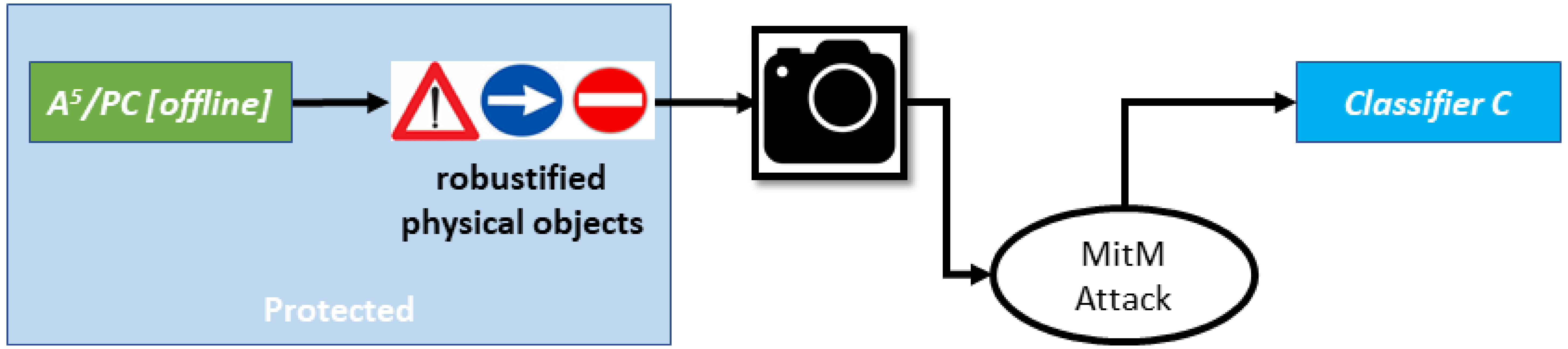
Results on CIFAR10, attack 8/255

Method	Clean error	Certified error
CROWN-IBP	54.02%	66.94%
A^5/O	45.68%	49.74%
A^5/R	50.91%	57.95%
A^5/RC	35.26%	42.76%



Configurations: Physical and Classifier (A^5/P and A^5/PC)

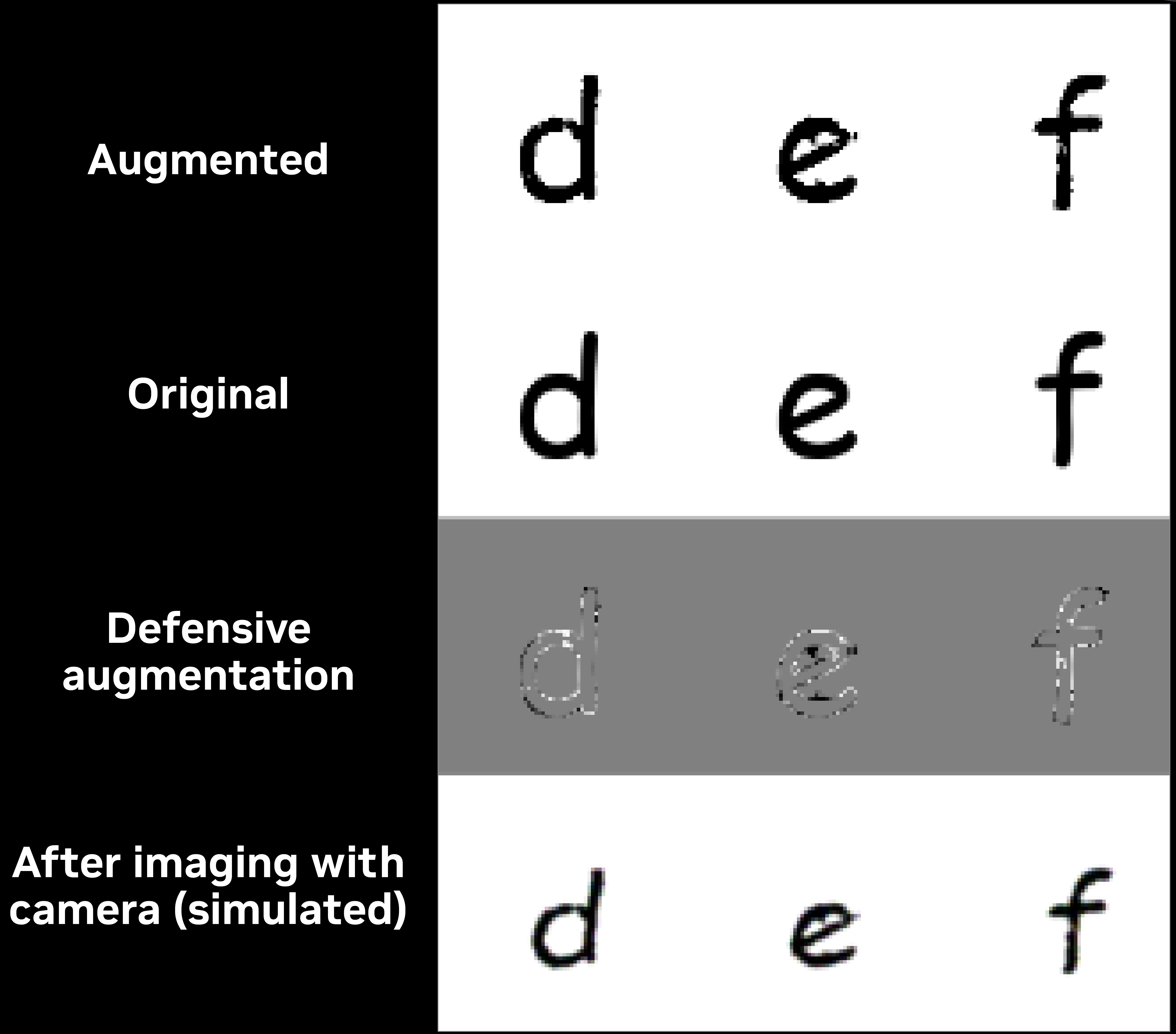
w
$w+\delta w_D$
A^5/PC
Protected
Acquisition A
$x+\delta x_D=A(w+\delta w_D)$
Protected
$x+\delta x_D+\delta x_A$
MitM Attack
Classifier C
A^5/PC
Protected



Configurations: Physical and Classifier (A^5/PC)

Results on OCR

Method	Clean error	Certified error
Vanilla	0.89%	100.00%
CROWN-IBP	3.85%	13.85%
A^5/P	3.08%	11.84%
A^5/PC	0.73%	4.20$



Augmented

Original

Defensive augmentation

After imaging with camera (simulated)

Conclusion

- Preemptive, certifiable robustification
- Offline and on the fly for acquired images
- Offline for physical objects
- The benefit of co-training
- Technical details and more results on MNIST, FashionMNIST, TinyImage net available in our paper
- Scaling to large network architecture still problematic
- Code available for research here: