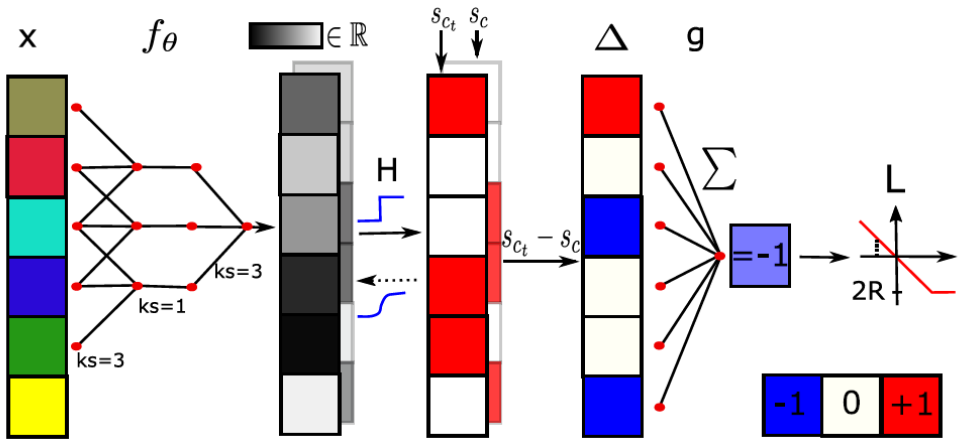
Efficient Certified Defenses Against Patch Attacks on Image Classifiers

Ninth International Conference on Learning Representations, ICLR 2021

Jan Hendrik Metzen

Maksym Yatsura



BOSCH

Invented for life

Motivation

- Adversarial patch threat model

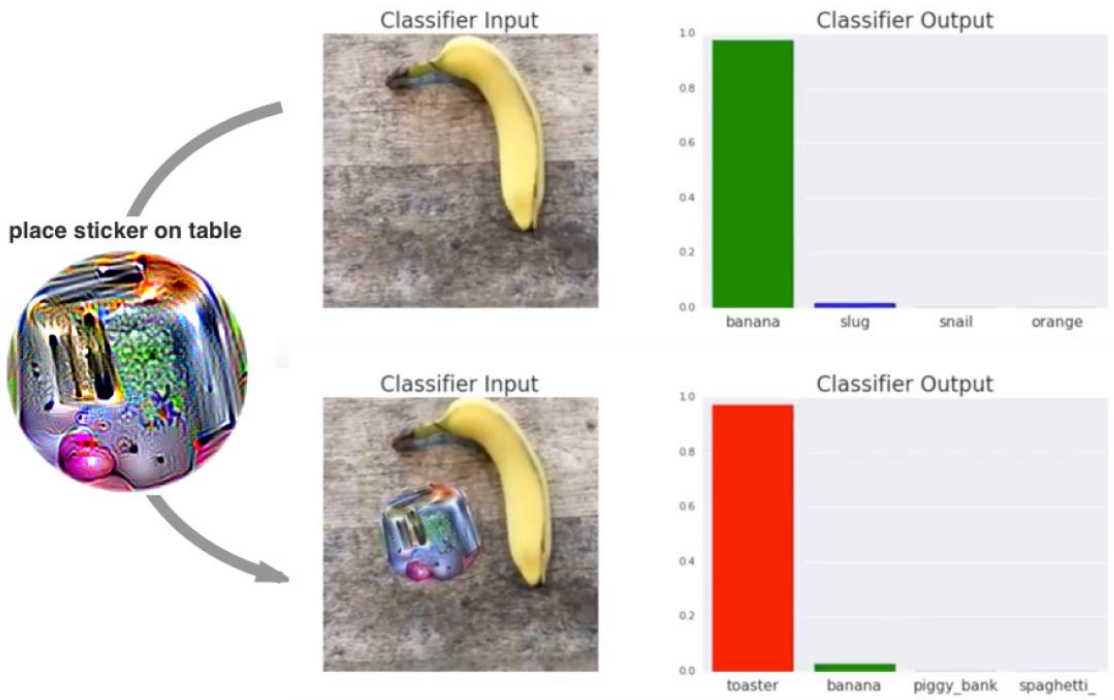


Image source: Brown et. al. *Adversarial Patch*, NeurIPS 2017

BOSCH

Motivation

- Adversarial patch threat model

- Physical world attacks on autonomous systems via their perception component

Image source: Metzen et. al. *Meta Adversarial Training*, 2021, https://arxiv.org/abs/2101.11453

BOSCH

Motivation

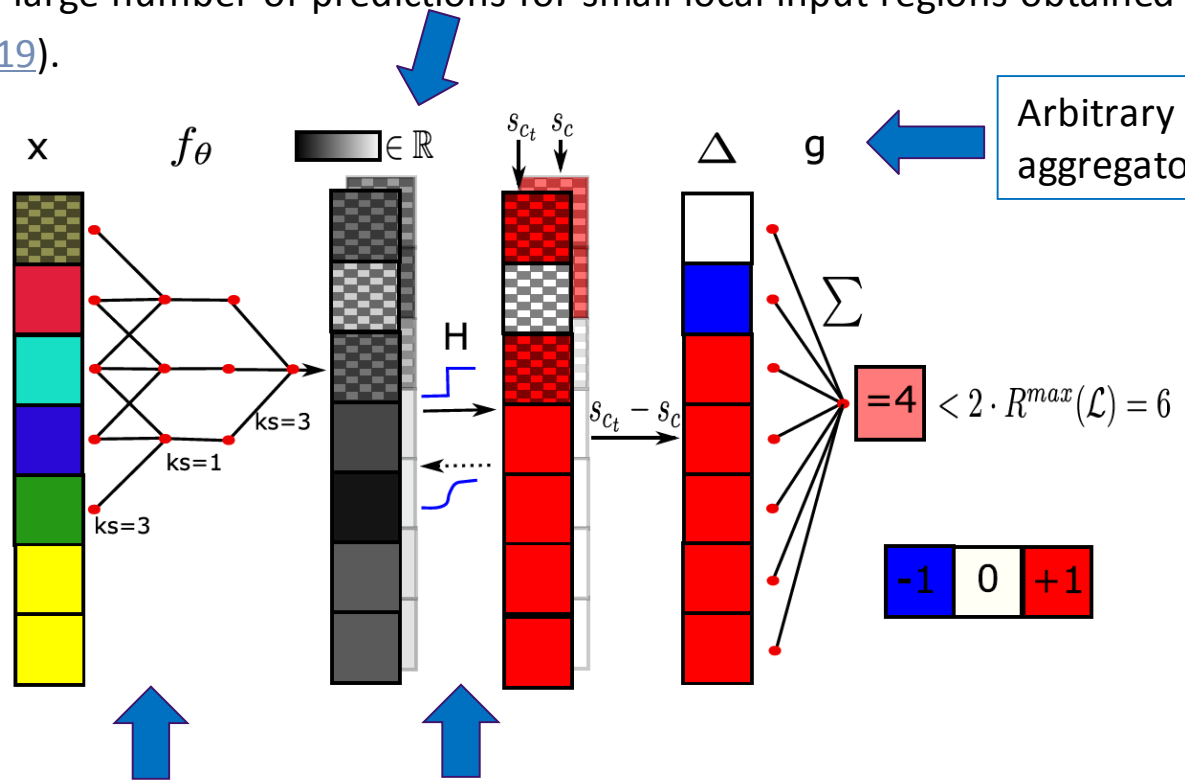
- Adversarial patch threat model
- Physical world attacks on autonomous systems via their perception component
- Safety-critical applications require a fail-safe fallback component with the following properties:
  - certifiable robustness against patch attacks
  - efficient inference
  - high performance on clean inputs

Image source: Metzen et. al. *Meta Adversarial Training*, 2021, https://arxiv.org/abs/2101.11453

BOSCH

BagCert: Architecture

Majority voting over a large number of predictions for small local input regions obtained with a BagNet-type architecture (Brendel & Bethge, 2019).



Arbitrary monotone spatial aggregator g can be used

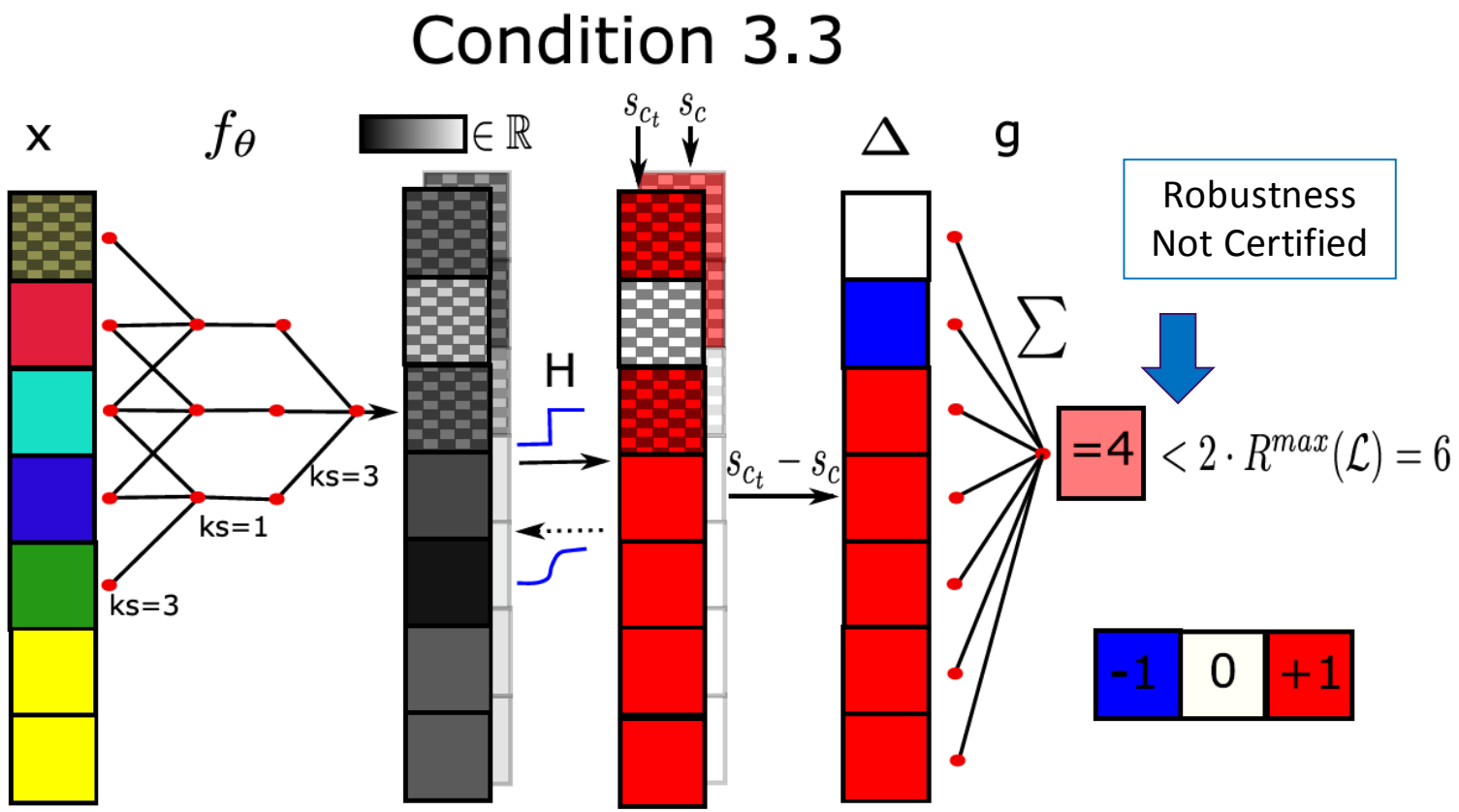
Small receptive field ensures that a local patch can only affect a small subset of predictions

Clipping via Heaviside function H(x) ensures that patches cannot increase individual scores arbitrarily

BOSCH

BagCert: Robustness Certification

Our Certification Condition 3.3 corresponds to the one by proposed [Levine & Feizi (2020)]

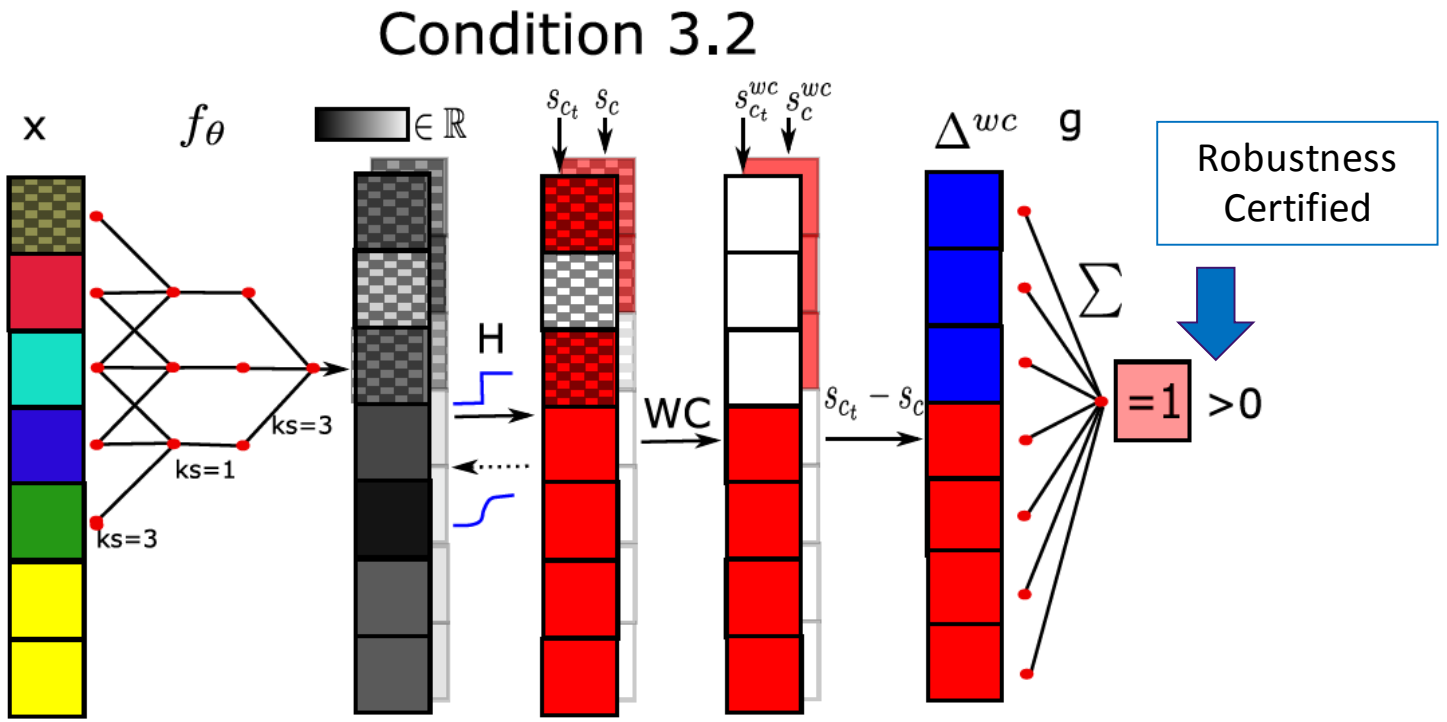


BOSCH

BagCert: Robustness Certification

Certification Condition 3.2 provides tighter bound: allows certifying robustness in more cases than previous work

- improves certified accuracy of the same model by roughly 3 percent points on CIFAR-10.
- preferable if all score regions affected by patches are rectangular or if there is only moderate number of possible patch locations.

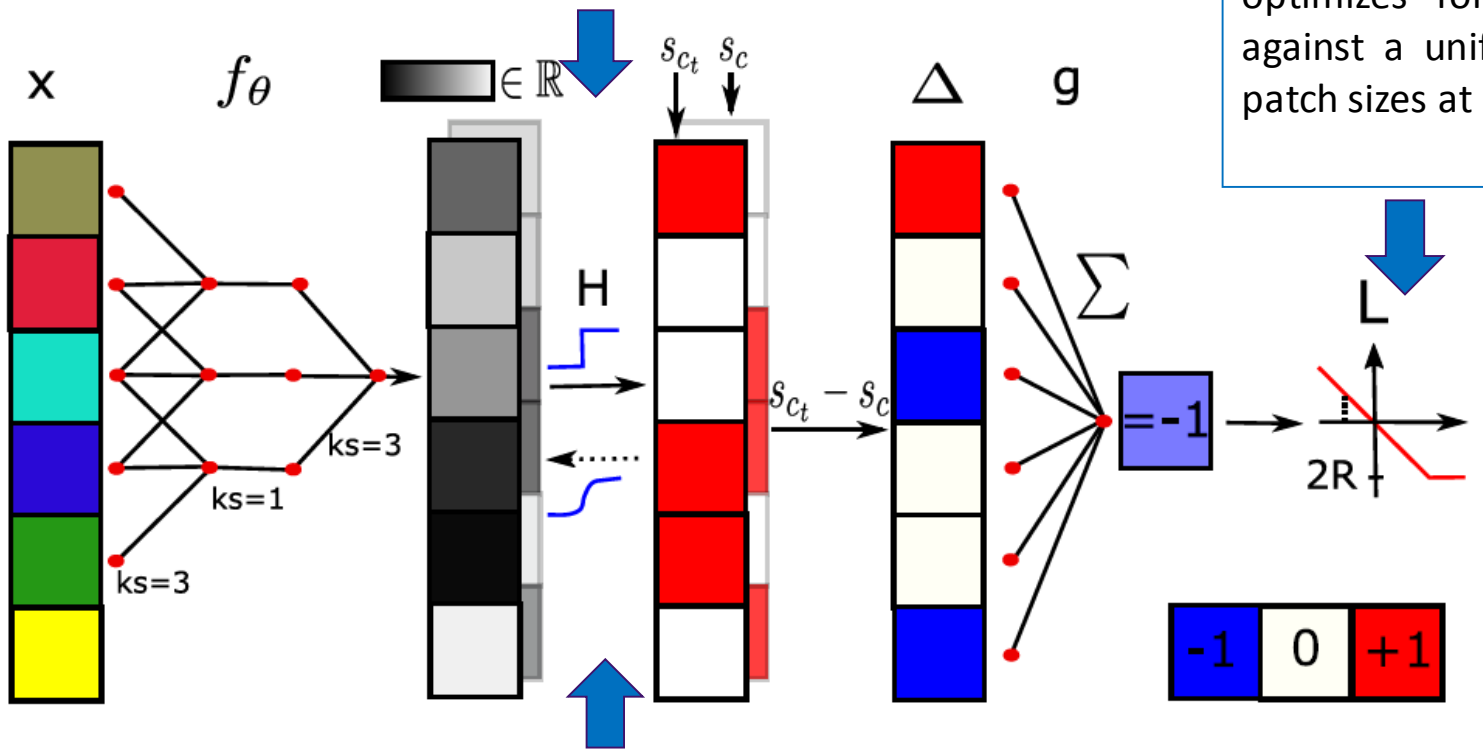


BOSCH

BagCert: End-to-end Training

End-to-end training obliterates need for threshold in clipping via H

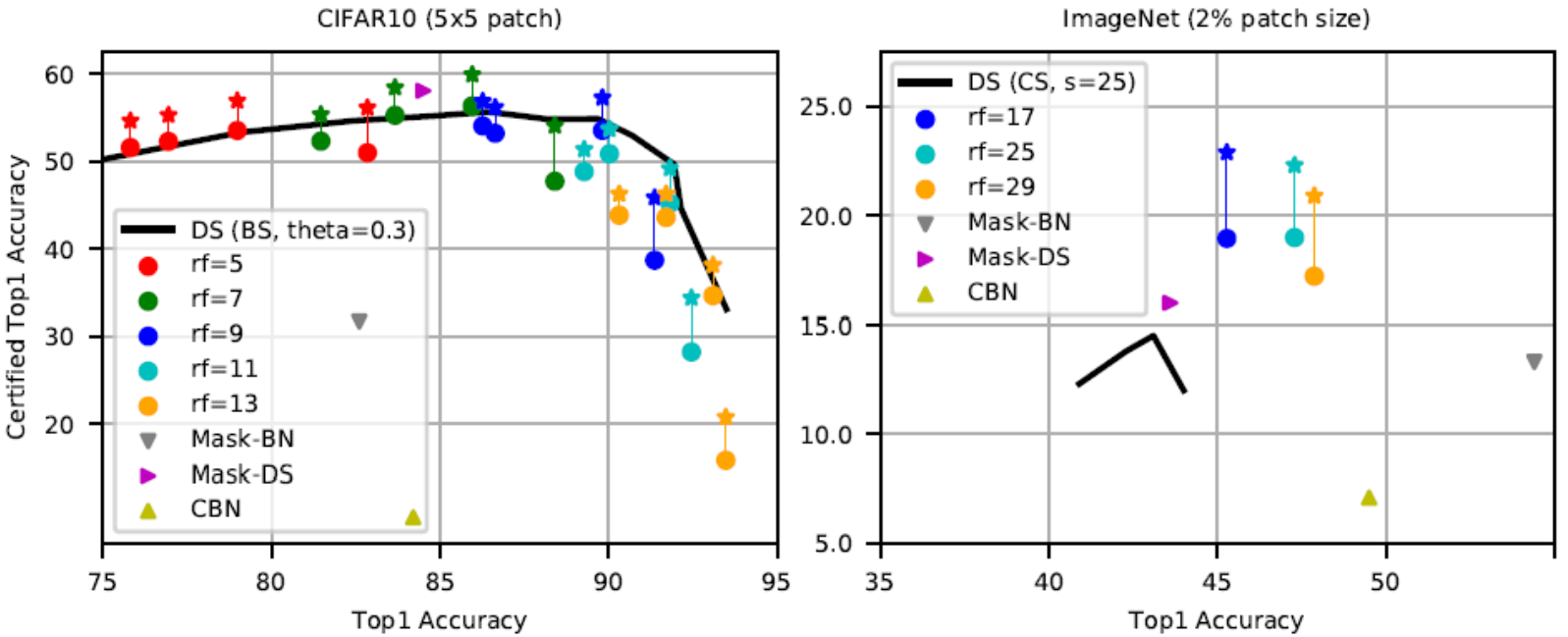
Margin-type loss function directly optimizes for certified accuracy against a uniform distribution of patch sizes at arbitrary positions



Backward pass smoothing via straight-through trick H -> sigmoid

BOSCH

Experimental Results: CIFAR10 and ImageNet



Plot legend:

Circles – certified via condition 3.3

Stars – certified via condition 3.2

DS – Derandomized smoothing
BS/CS – batch/column smoothing
(Levine & Feizi, 2020)

Mask-BN – Masked BagNet
(Xiang et. al., 2020)

Mask-DS – Masked DS-ResNet
(Xiang et. al., 2020)

CBN – Clipped BagNet
(Zhang et. al., 2020)

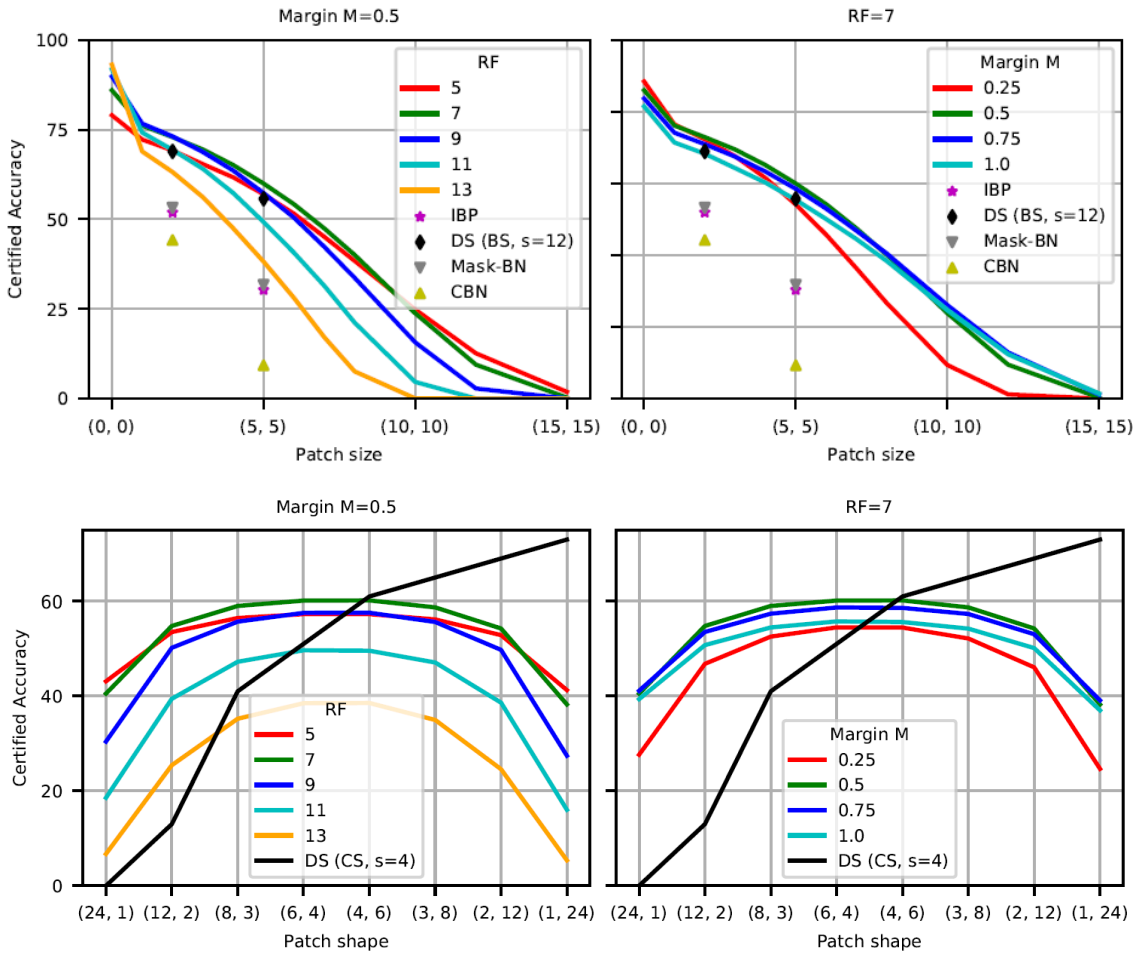
RF of BagCert	5	7	9	11	13	DS(BS)	DS(CS)
Cert. time (10.000 images)	39.0s	40.6s	43.2s	45.9s	48.5s	**788.0s**	28.0s
Number of parameters	28M	38M	57M	57M	66M	11M	11M

BOSCH

Experimental Results: Patch Sizes and Aspect Ratios

- A single configuration of BagCert (receptive field size 7, margin $M = 0.5$) performs close to optimal for all patch sizes
  - can certify non-trivial performance for up to 10×10 patch size.

- BagCert has stable performance over different patch shapes unlike column smoothing proposed by Levine & Feizi, 2020.



BOSCH

BagCert and Derandomized smoothing (Levine & Feizi, 2020)

	Block Smoothing	Column/row Smoothing	BagCert
Efficient certification	No	Yes	Yes
Robustness against non-square patches	Yes	No	Yes

For more technical details, please visit our poster.

We are looking forward to the discussion!

BOSCH