Scalable Certified Segmentation via Randomized Smoothing

Marc
Fischer

Maximillian
Baader

Martin
Vechev

ETH zürich

safeai.ethz.ch

SRILAB

Adversarial Attack for Segmentation

Randomized Smoothing [Cohen et al.]

$$\bar{f}(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}} \, \mathbb{P}(f(x+\epsilon) = c)$$

for classifier f, noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$

Then $\bar{f}(x)$ = $\bar{f}(x+\delta)$ for $\|\delta\|_2 \leq R$.

Randomized Smoothing [Cohen et al.]

$$\bar{f}(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}(f(x+\epsilon) = c)$$

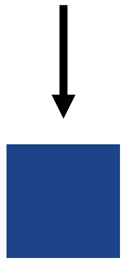
for classifier f, noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$
Then $\bar{f}(x) = \bar{f}(x+\delta)$ for $\|\delta\|_2 \leq R$.

function CERTIFY($f, \sigma, \boldsymbol{x}, n_0, n, \alpha$)
  $\texttt{cnts}^0 \leftarrow \text{SAMPLE}(f, \boldsymbol{x}, n_0, \sigma)$
  $\hat{c}_A \leftarrow$ top index in $\texttt{cnts}^0$
  $\texttt{cnts} \leftarrow \text{SAMPLE}(f, \boldsymbol{x}, n, \sigma)$
  $\underline{p_A} \leftarrow \text{LOWERCONFBND}(\texttt{cnts}[\hat{c}_A], n, 1-\alpha)$
  if $\underline{p_A} > \frac{1}{2}$ **return** prediction $\hat{c}_A$ and radius $\sigma\, \Phi^{-1}(\underline{p_A})$
  else return $\oslash$

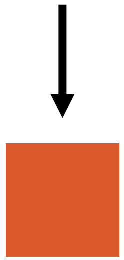
In pratice, approximated via sampling:
$\bar{f}(x) = \bar{f}(x+\delta)$ for $\|\delta\|_2 \leq \sigma\Phi^{-1}(\underline{p_A})$ with confidence $1-\alpha$.

Classification

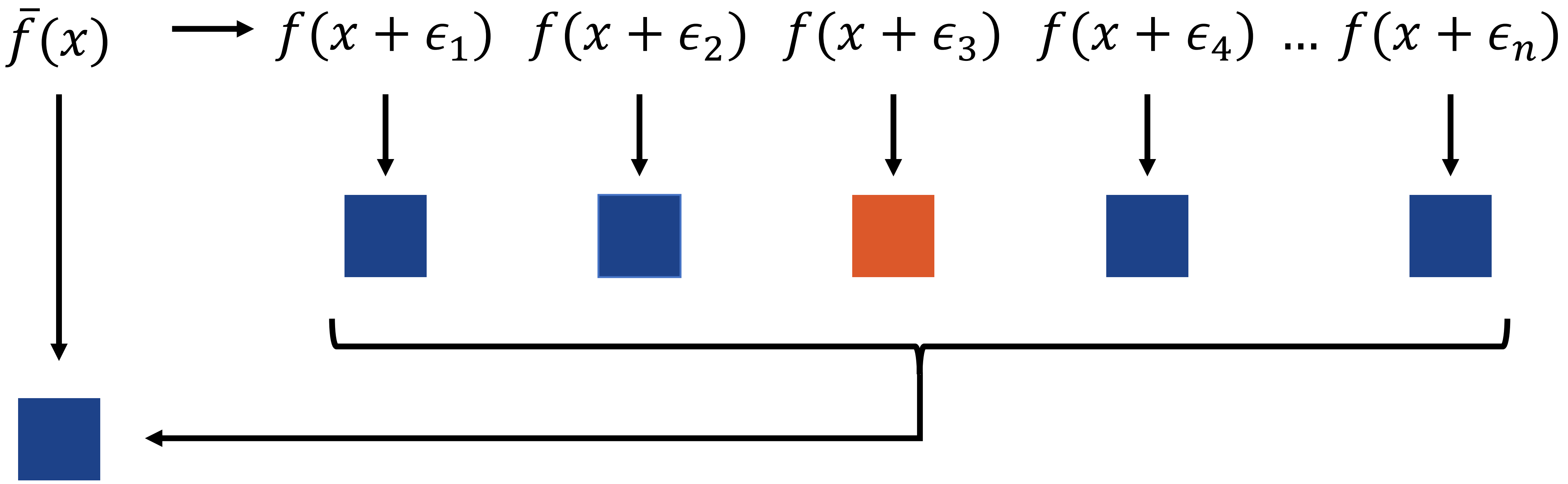
$f(x)$



$f(x + \delta)$

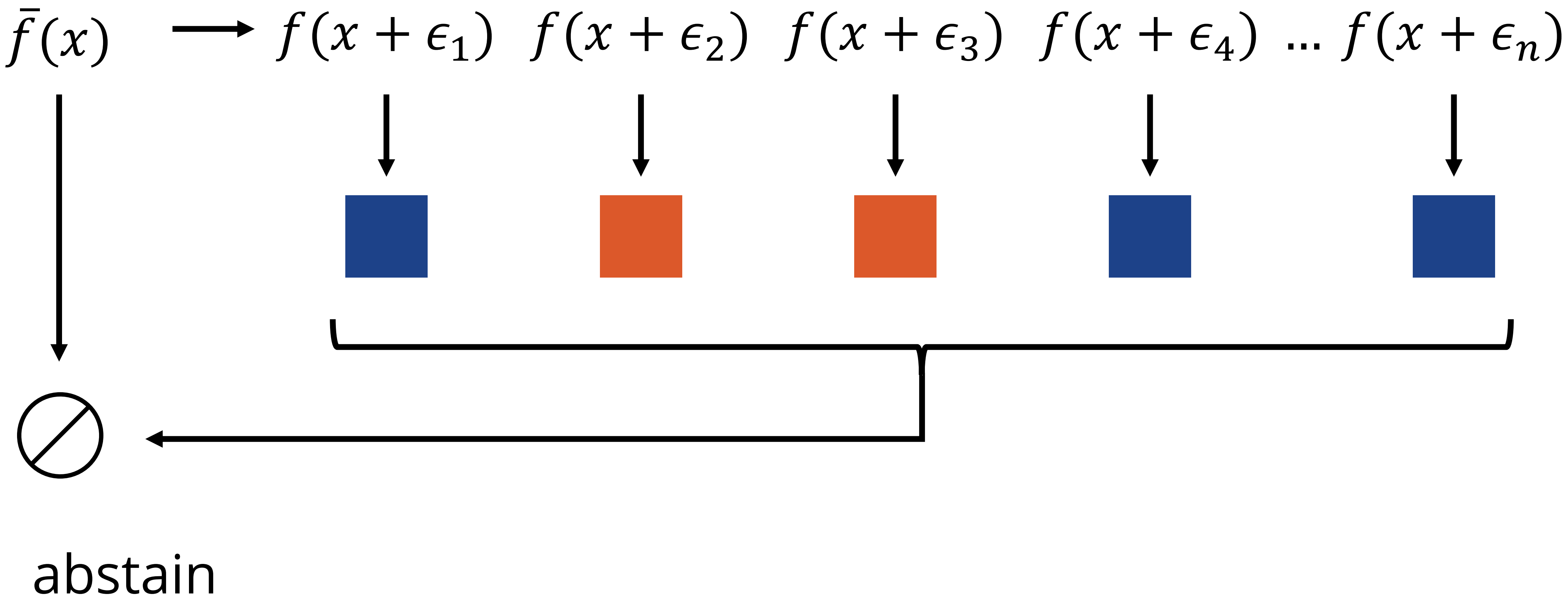


Randomized Smoothing [Cohen et al.]



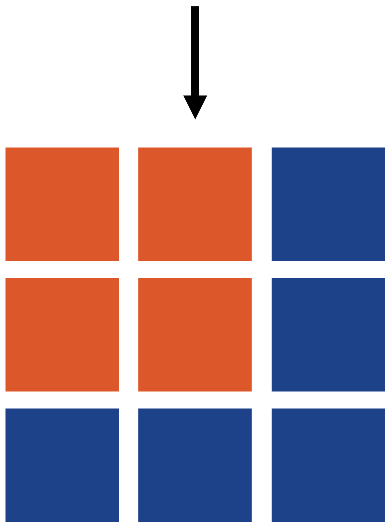
robustness radius R
with confidence $1 - \alpha$

Randomized Smoothing [Cohen et al.]

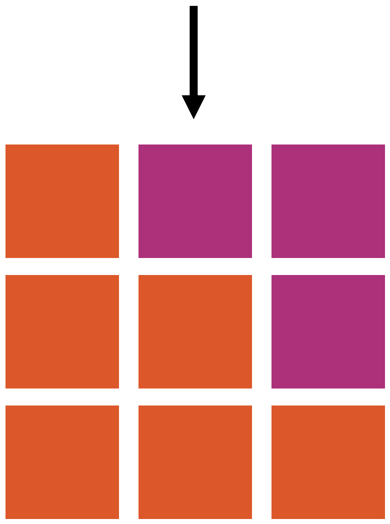


Segmentation

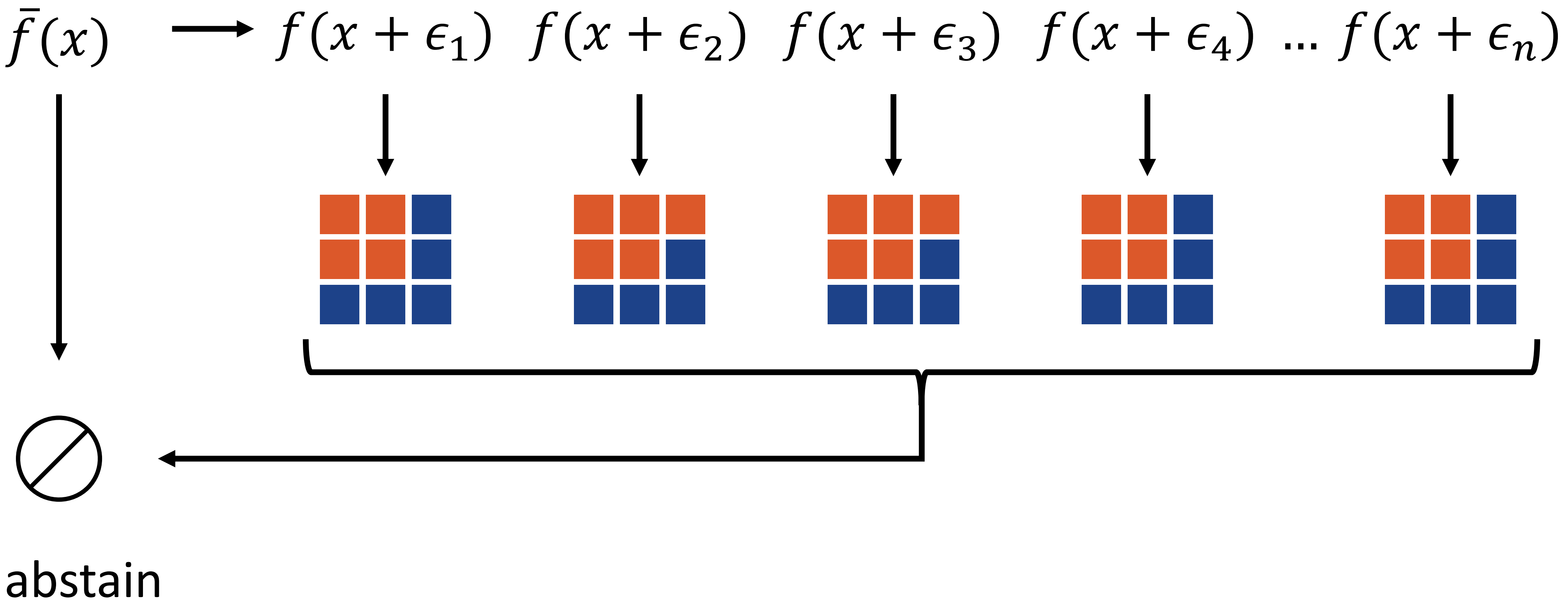
$f(x)$



$f(x + \delta)$



Naïve Randomized Smoothing for Segmentation



Key Challenges

Bad Components: a single component that is unstable under noise, can cause abstention or dominate radius R

Multiple Testing: as individual results only hold w.h.p, obtaining high overall confidence is challenging

Randomized Smoothing for Segmentation

$$\bar{f}_i^{\tau}(x) = \begin{cases} c & \text{if } \mathbb{P}(f_i(x+\epsilon) = c) > \tau \\ \emptyset & \text{else} \end{cases}$$

for segmentation model f,
noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$

Randomized Smoothing for Segmentation

$$\bar{f}_i^{\tau}(x) = \begin{cases} c & \text{if } \mathbb{P}(f_i(x+\epsilon) = c) > \tau \\ \emptyset & \text{else} \end{cases}$$

for segmentation model f,
noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$

Then $\bar{f}_i^{\tau}(x) = \bar{f}_i^{\tau}(x+\delta)$, $i \in I_x \coloneqq \{i \mid \bar{f}_i^{\tau}(x) \neq \emptyset\}$
for $\|\delta\|_2 \leq R \coloneqq \sigma\Phi^{-1}(\tau)$.

Randomized Smoothing for Segmentation

$$\bar{f}_i^{\tau}(x) = \begin{cases} c & \text{if } \mathbb{P}(f_i(x+\epsilon) = c) > \tau \\ \emptyset & \text{else} \end{cases}$$

for segmentation model f,
noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$

Then $\bar{f}_i^{\tau}(x) = \bar{f}_i^{\tau}(x+\delta)$, $i \in I_x \coloneqq \{i \mid \bar{f}_i^{\tau}(x) \neq \emptyset\}$
for $\|\delta\|_2 \leq R \coloneqq \sigma\Phi^{-1}(\tau)$.

In pratice, via sampling obtain $\hat{I}_x$ s.t. with confidence $1-\alpha$, $\hat{I}_x \subseteq I_x$.

function SEGCERTIFY($f, \sigma, \boldsymbol{x}, n, n_0, \tau, \alpha$)
    $\texttt{cnts}_1^0, \ldots, \texttt{cnts}_N^0 \leftarrow$ SAMPLE($f, \boldsymbol{x}, n_0, \sigma$)
    $\texttt{cnts}_1, \ldots, \texttt{cnts}_N \leftarrow$ SAMPLE($f, \boldsymbol{x}, n, \sigma$)
    for $i \leftarrow \{1, \ldots, N\}$:
        $\hat{c}_i \leftarrow$ top index in $\texttt{cnts}_i^0$
        $n_i \leftarrow \texttt{cnts}_i[\hat{c}_i]$
        $pv_i \leftarrow$ BINPVALUE($n_i, n, \leq, \tau$)
    $r_1, \ldots, r_N \leftarrow$ FWERCONTROL($\alpha, pv_1, \ldots, pv_N$)
    for $i \leftarrow \{1, \ldots, N\}$:
        if $\neg r_i$: $\hat{c}_i \leftarrow \emptyset$
    $R \leftarrow \sigma\Phi^{-1}(\tau)$
    return $\hat{c}_1, \ldots, \hat{c}_N, R$

Randomized Smoothing for Segmentation

$$\bar{f}_i^{\tau}(x) = \begin{cases} c & \text{if } \mathbb{P}(f_i(x+\epsilon) = c) > \tau \\ \emptyset & \text{else} \end{cases}$$

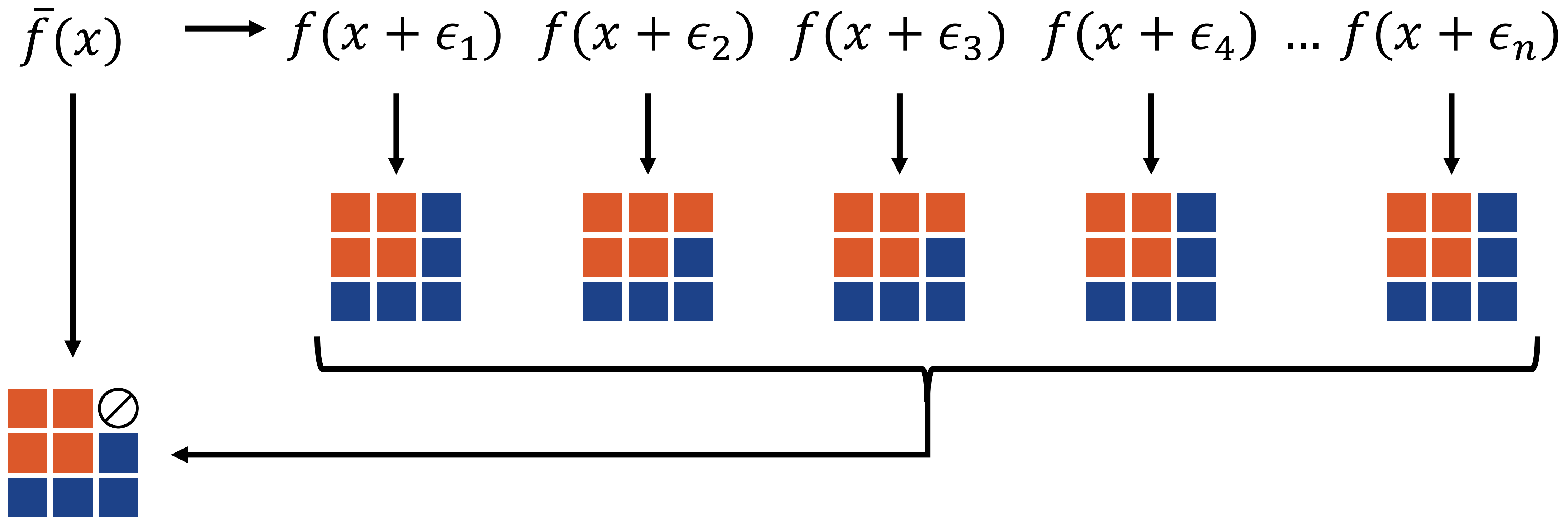
for segmentation model f,
noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$

function SEGCERTIFY$(f, \sigma, \boldsymbol{x}, n, n_0, \tau, \alpha)$
  $\texttt{cnts}_1^0, \ldots, \texttt{cnts}_N^0 \leftarrow$ SAMPLE$(f, \boldsymbol{x}, n_0, \sigma)$
  $\texttt{cnts}_1, \ldots, \texttt{cnts}_N \leftarrow$ SAMPLE$(f, \boldsymbol{x}, n, \sigma)$
  for $i \leftarrow \{1, \ldots, N\}$:
    $\hat{c}_i \leftarrow$ top index in $\texttt{cnts}_i^0$
    $n_i \leftarrow \texttt{cnts}_i[\hat{c}_i]$
    $pv_i \leftarrow$ BINPVALUE$(n_i, n, \leq, \tau)$
  $r_1, \ldots, r_N \leftarrow$ FWERCONTROL$(\alpha, pv_1, \ldots, pv_N)$
  for $i \leftarrow \{1, \ldots, N\}$:
    if $\neg r_i$: $\hat{c}_i \leftarrow \emptyset$
  $R \leftarrow \sigma \Phi^{-1}(\tau)$
  return $\hat{c}_1, \ldots, \hat{c}_N, R$

Then $\bar{f}_i^{\tau}(x) = \bar{f}_i^{\tau}(x+\delta), i \in I_x \coloneqq \{i \mid \bar{f}_i^{\tau}(x) \neq \emptyset\}$
for $\|\delta\|_2 \leq R \coloneqq \sigma\Phi^{-1}(\tau)$.

In pratice, via sampling obtain $\hat{I}_x$ s.t. with confidence $1 - \alpha$, $\hat{I}_x \subseteq I_x$.

Randomized Smoothing for Segmentation



robustness radius R
with confidence $1 - \alpha$

Semantic Segmentation

	cert radius	pixel acc.	mIoU	abstain
non-robust	-	0.96	0.76	-
base model	-	0.89	0.51	-
certified	0.34	0.86	0.54	0.10

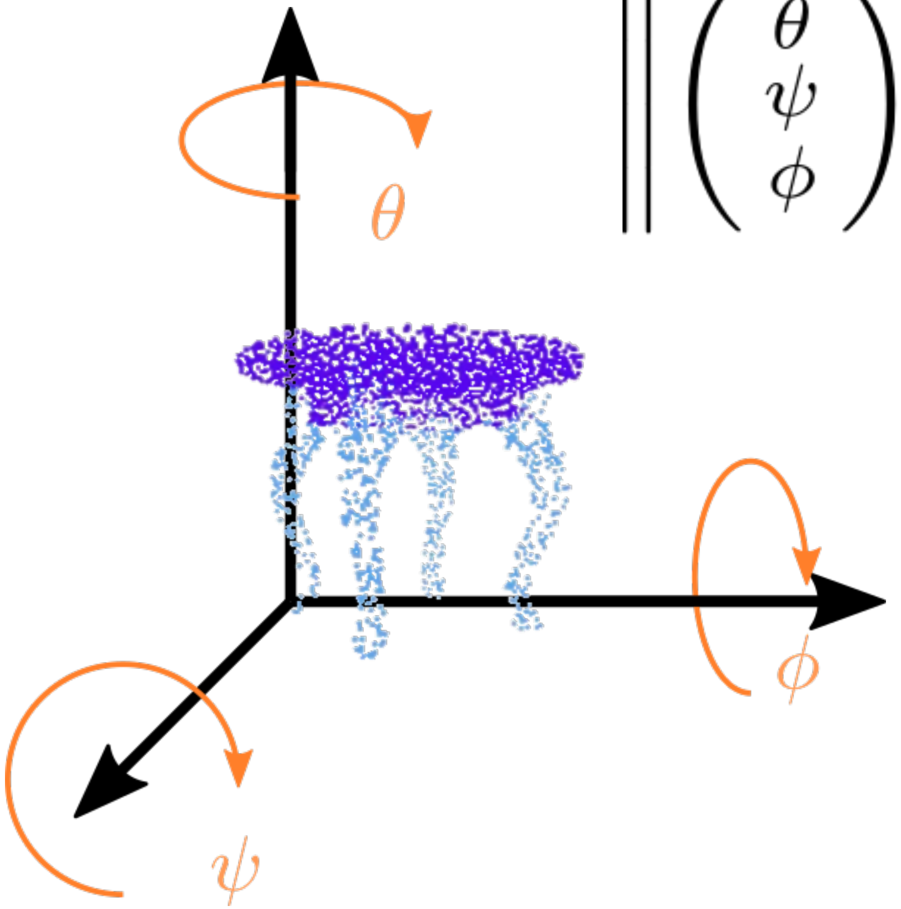
HrNetV2 on Cityscapes evaluated on 100 images,
$\sigma = 0.5$, $n = 100$ samples, scale 0.5

Point Cloud Part Segmentation

	cert radius	pixel acc.	abstain
non-robust	-	0.91	-
base model	-	0.86	-
certified	0.26	0.71	0.25

PointNetV2 on ShapeNet evaluated on 100 inputs,
$\sigma = 0.25$, $n = 1000$ samples

Point Cloud Part Segmentation, Rotation

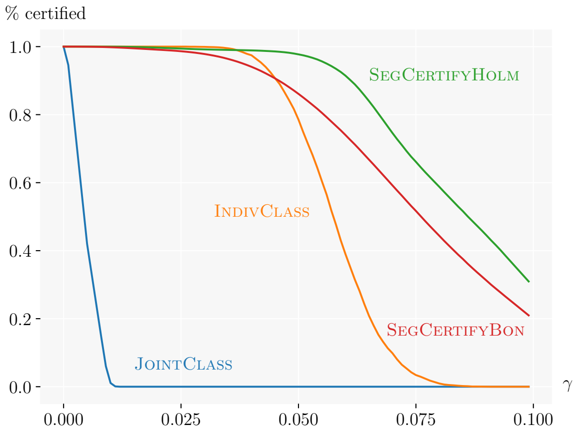


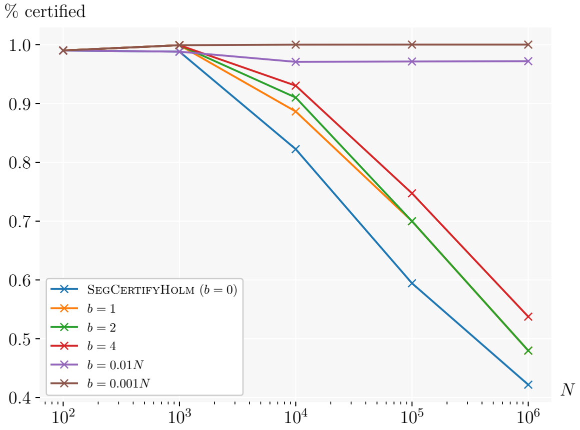
$$\left\| \begin{pmatrix} \theta \\ \psi \\ \phi \end{pmatrix} \right\|_2 \leq R$$

	cert radius	pixel acc.	abstain
non-robust	-	0.91	-
base model	-	0.77	-
certified	0.26	0.69	0.16

PointNetV2 on ShapeNet evaluated on 100 inputs, $\sigma = 0.125$, $n = 1000$ samples

In the paper





- Motiation & Derivation
- Further results
- Effect of different FWER schemes
- Allowing error budgets