# PatchCURE: Improving Certifiable Robustness, Model Utility, and Computation Efficiency of Adversarial Patch Defenses

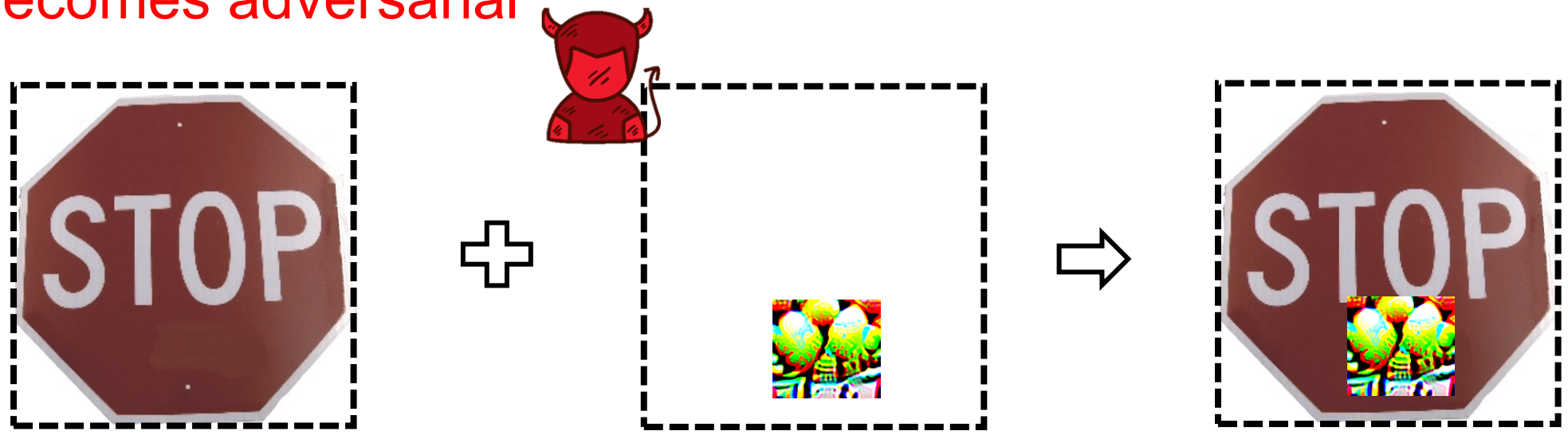
Chong Xiang[1], Tong Wu[1], Sihui Dai[1], Jonathan Petit[2], Suman Jana[3], Prateek Mittal[1]

*[1]Princeton University, [2]Qualcomm Technologies, Inc., [3]Columbia University*



# Adversarial Patch: A Variant of Adversarial Examples

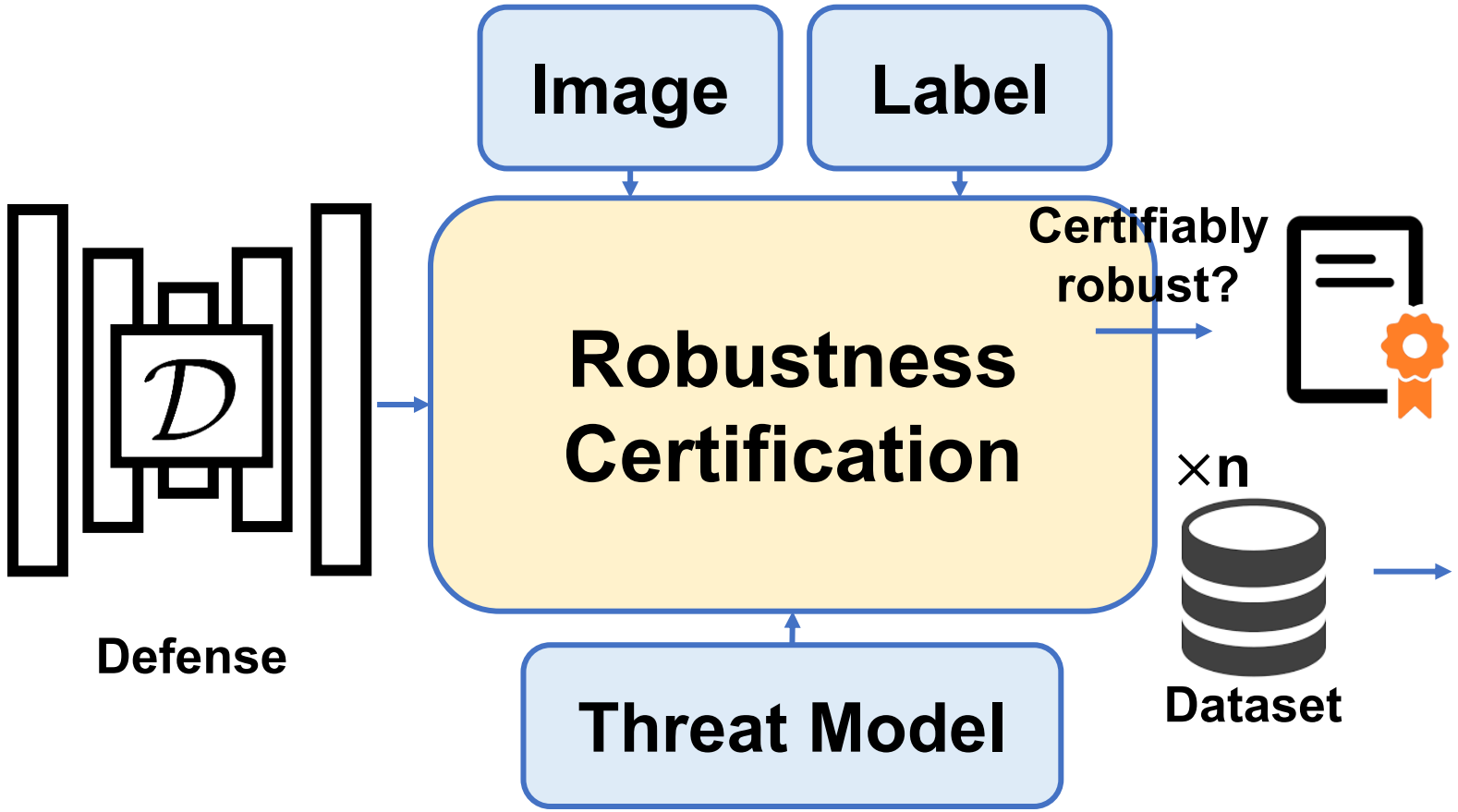
- The attacker has arbitrary control over pixels within a localized image region (i.e., a patch region)
- Optimize the patch content to induce misclassification
- Print and attach the patch – images taken from that physical scene becomes adversarial




**“Stop”**

**“Speed Limit 80”**

# Defense Objective: Certifiable Evaluation of Robustness



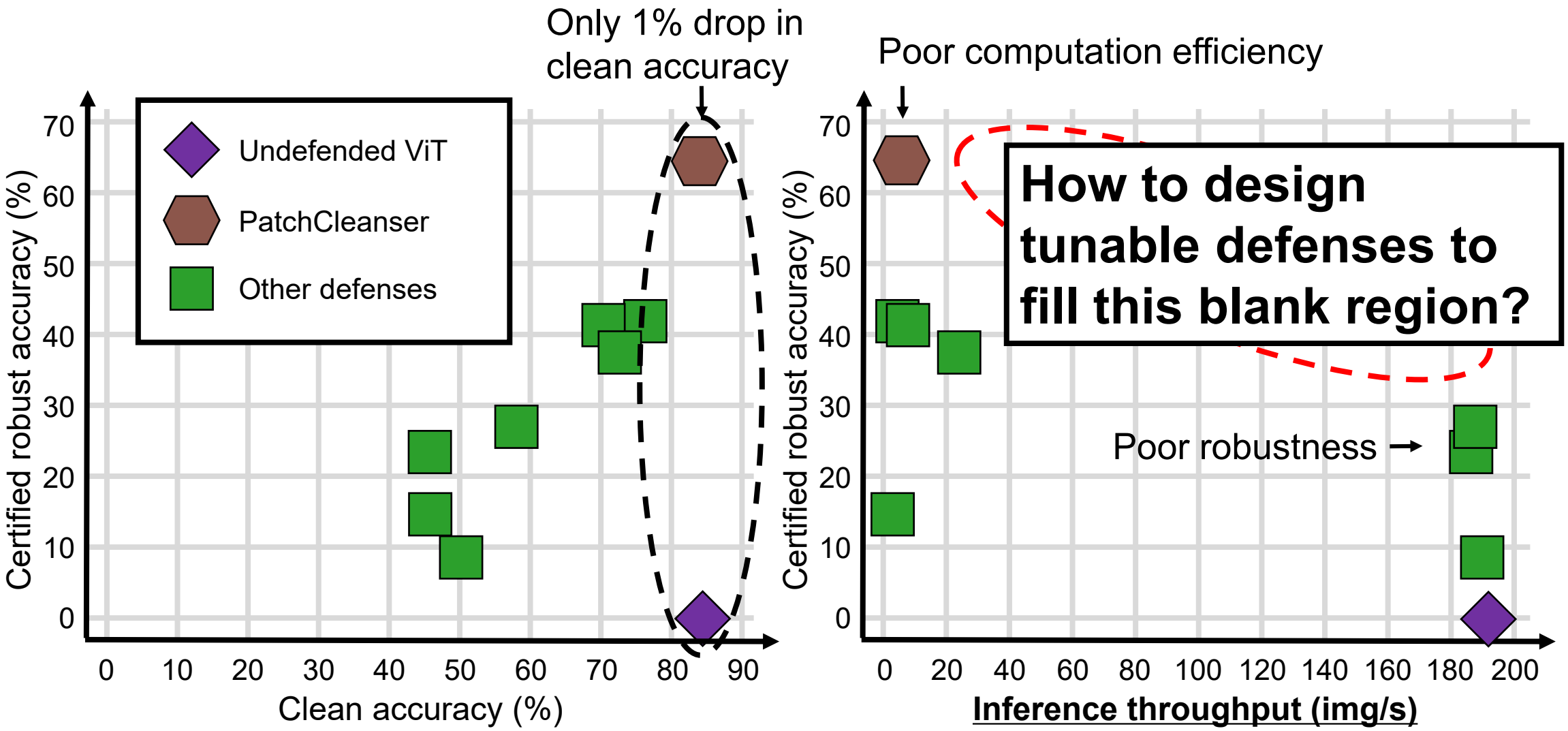

The model prediction on this image is *always* correct, no matter what a **white-box adaptive** attacker within the threat model does

**Certified Robust Accuracy:** The fraction of images with robustness certificate

**A provable lower bound!** (won't be compromised in the future)

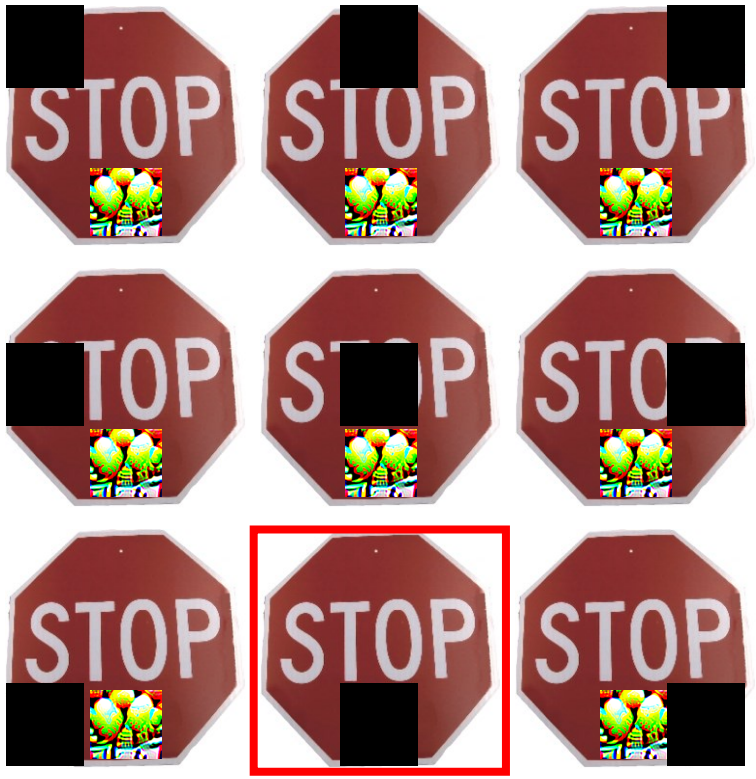
Example: a $32 \times 32$ patch on a $224 \times 224$ image, at any image location ($193^2$ possible cases), with any patch content ($2^{24576}$ possible cases)

# State of Research: Performance on ImageNet

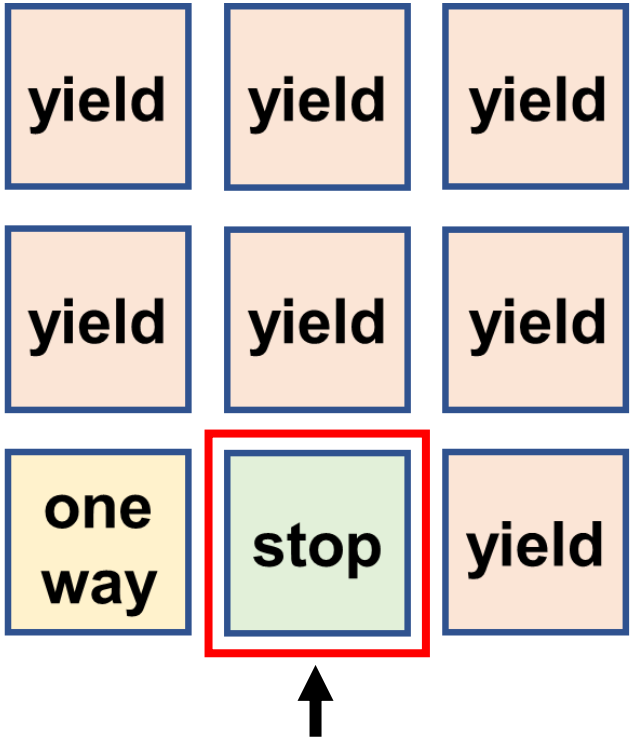


# Pixel Masking: A Powerful but Computation-intensive Defense

- Insight: when the corruption is localized, we can use a mask to remove it




**Masked images**

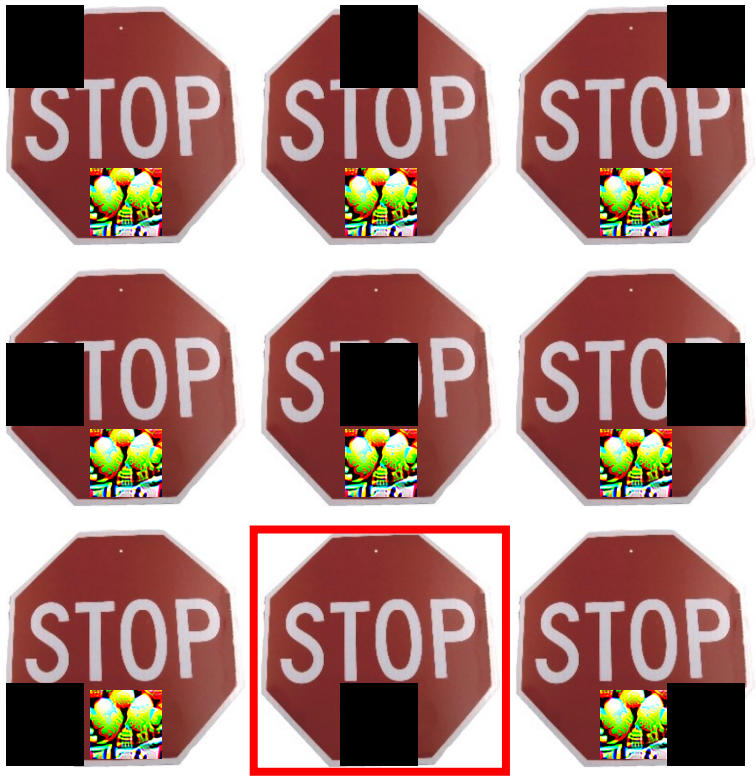



Correct label recovered when mask removes the entire patch

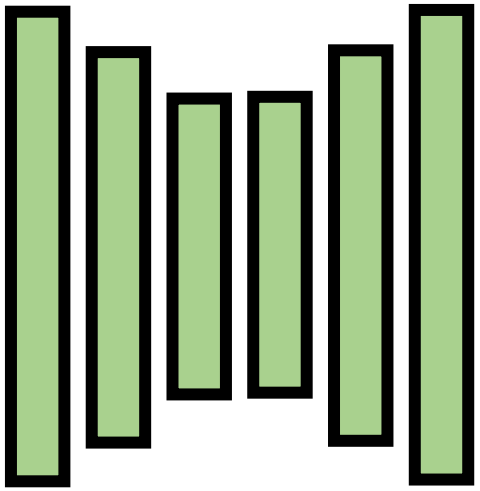
Analyze the **pattern of masked predictions** for a robust prediction

# Pixel Masking: A Powerful but Computation-intensive Defense

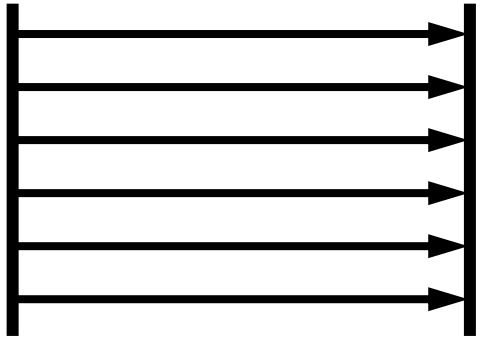
- Insight: when the corruption is localized, we can use a mask to remove it




**Masked images**



**Model**

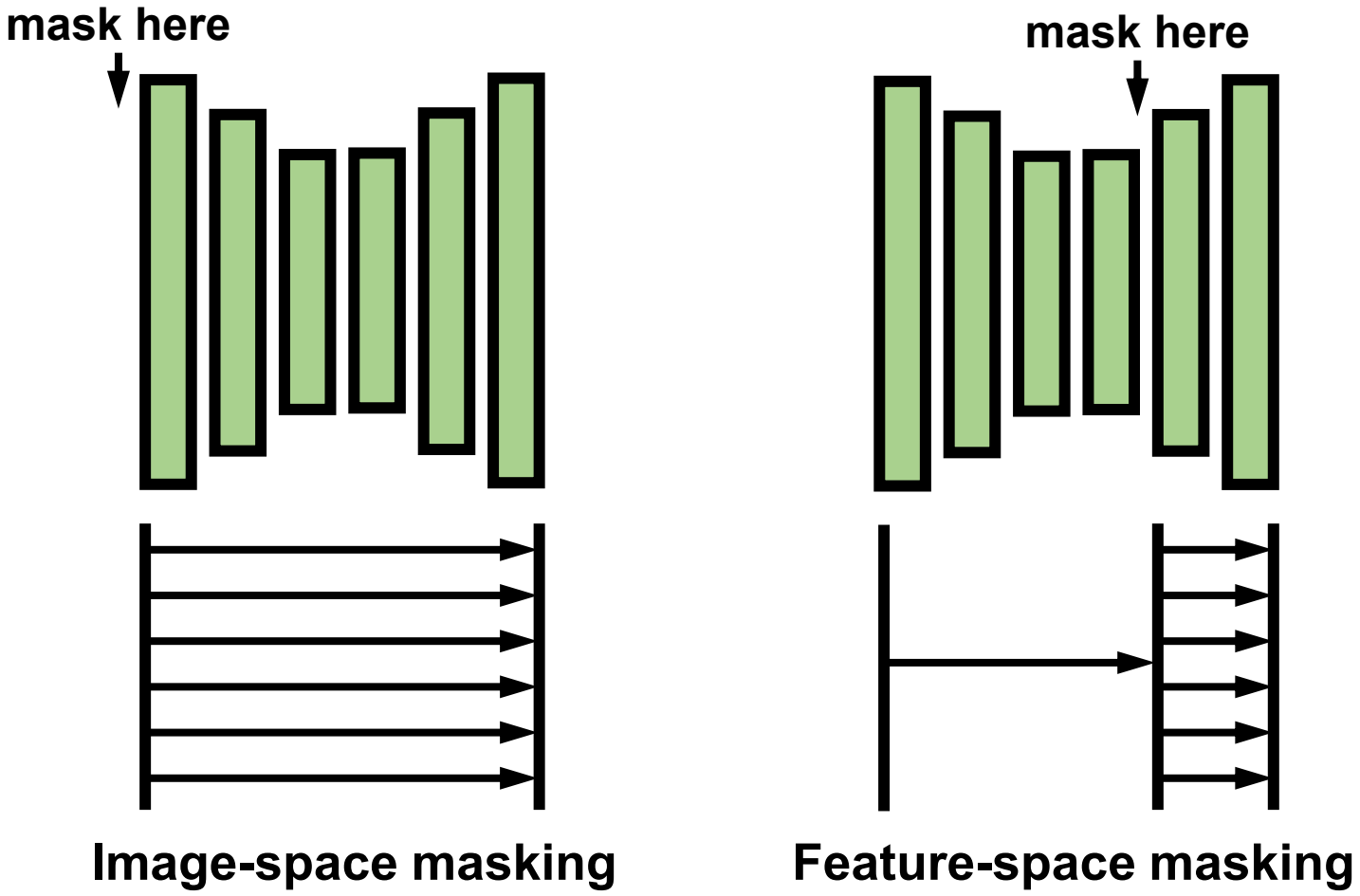


**Requires multiple (10-100x) end-to-end model feedforward passes**

**How to reduce computation overhead (without losing all certifiable robustness)?**

# PatchCURE Idea: Feature-space Masking

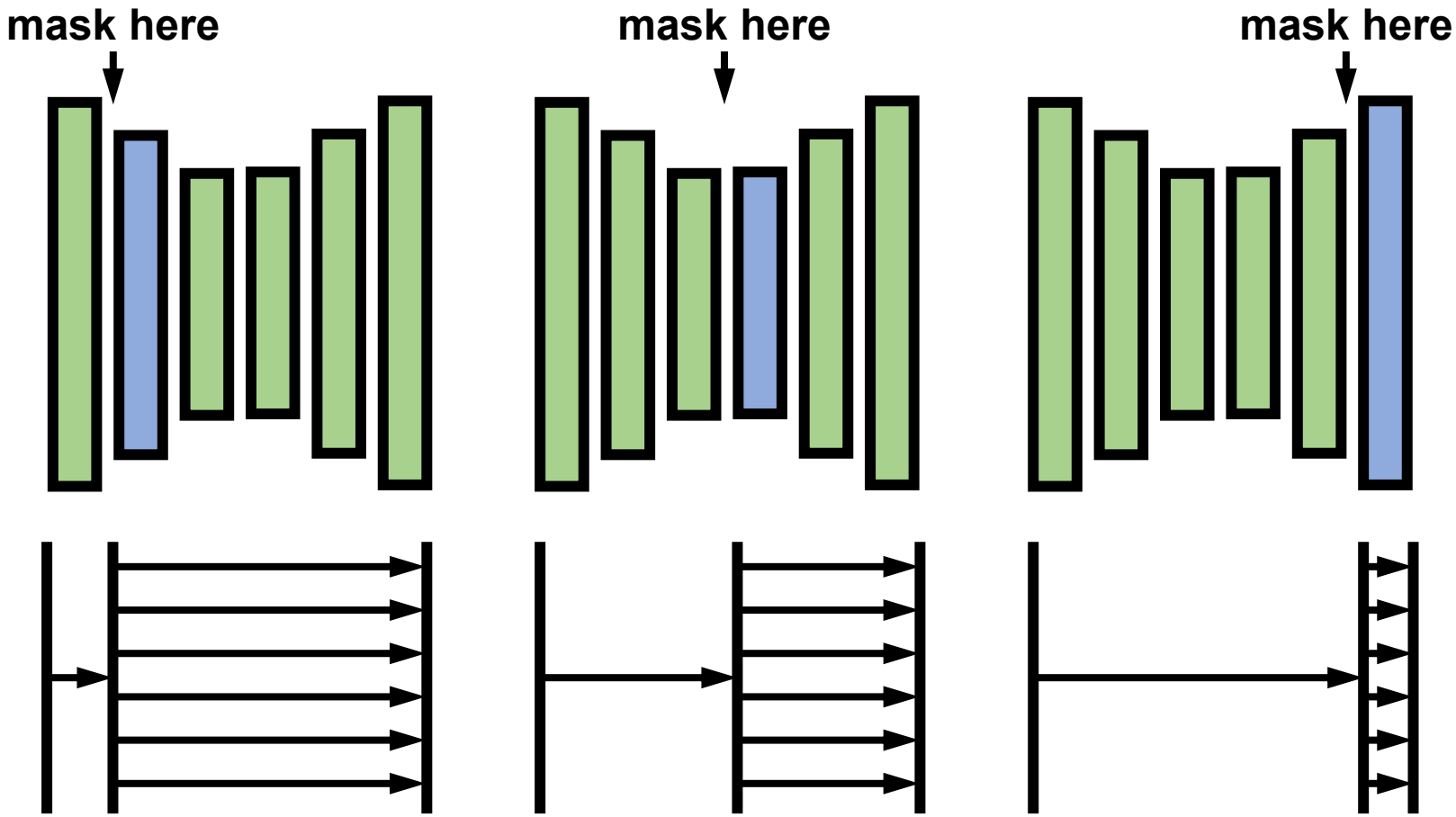
- **Insight: Only layers after the masking require repeated computation**




**Reuse the computation of layers before the masking!**

# PatchCURE Idea: Feature-space Masking

- **Insight: Only layers after the masking require repeated computation**

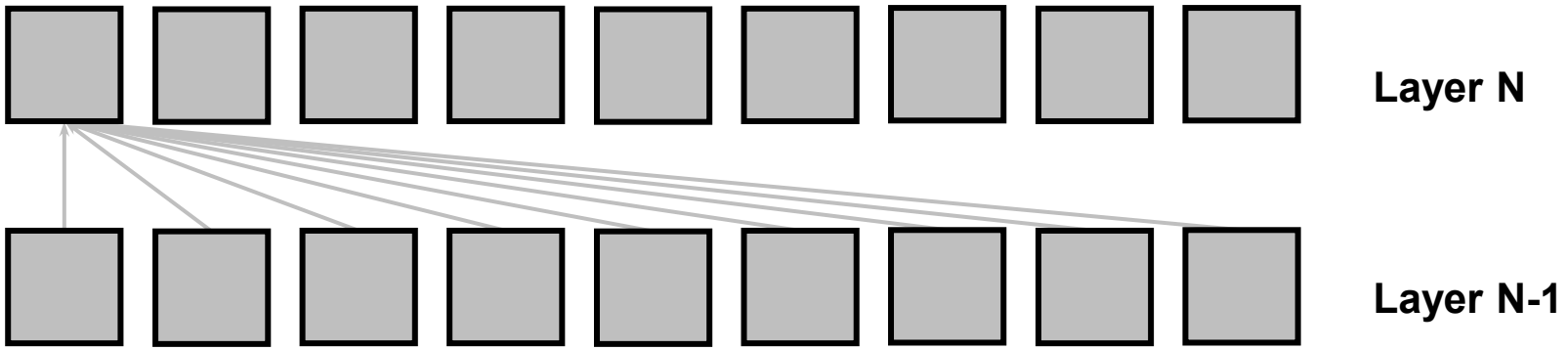



**Adjusting masking locations can systematically tune efficiency**

**What about robustness?**

# Naïve Feature-masking is not Robust

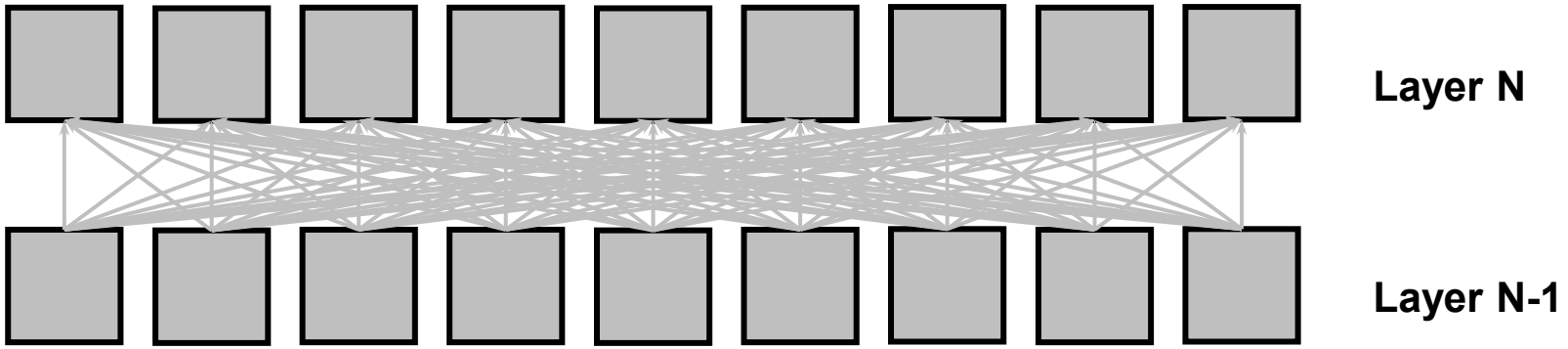
- Example: Global attention in Vision Transformer




- Every token/feature receives signals from all tokens in the previous layer (global receptive field)

# Naïve Feature-masking is not Robust

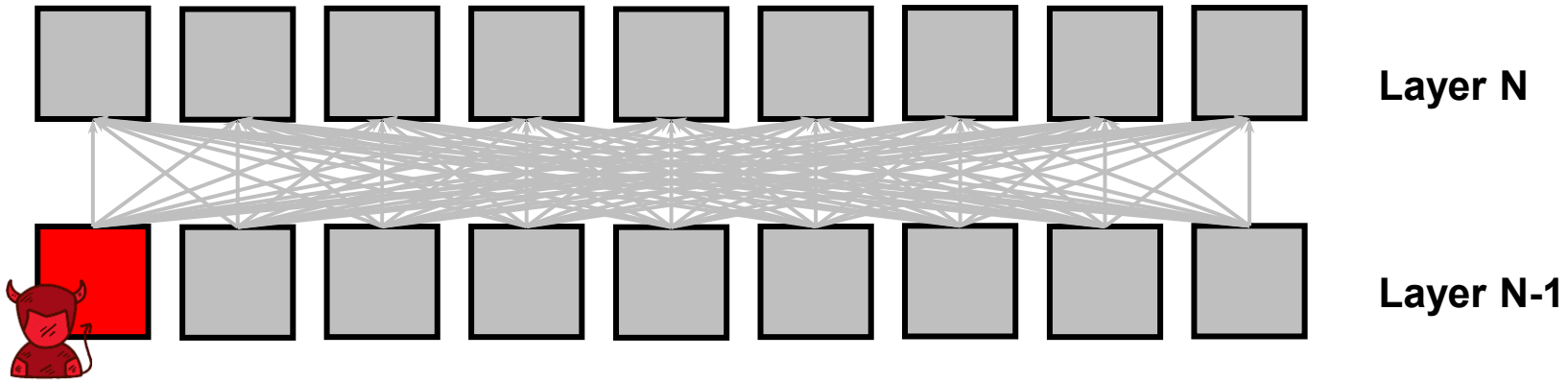
- Example: Global attention in Vision Transformer




- Every token/feature receives signals from all tokens in the previous layer (global receptive field)

# Naïve Feature-masking is not Robust

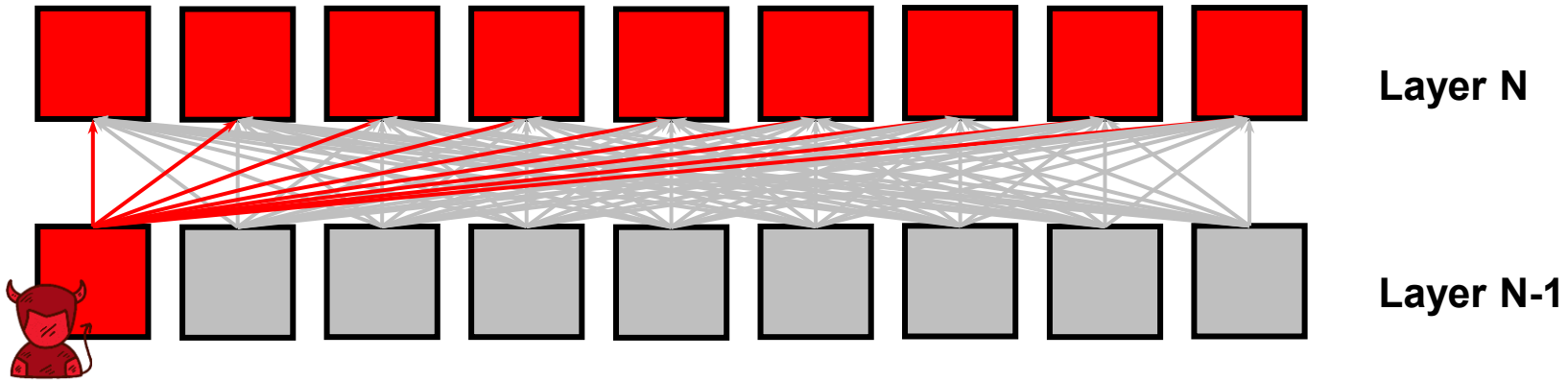
- Example: Global attention in Vision Transformer




- Every token/feature receives signals from all tokens in the previous layer (global receptive field)
- **Large receptive fields can hurt robustness**

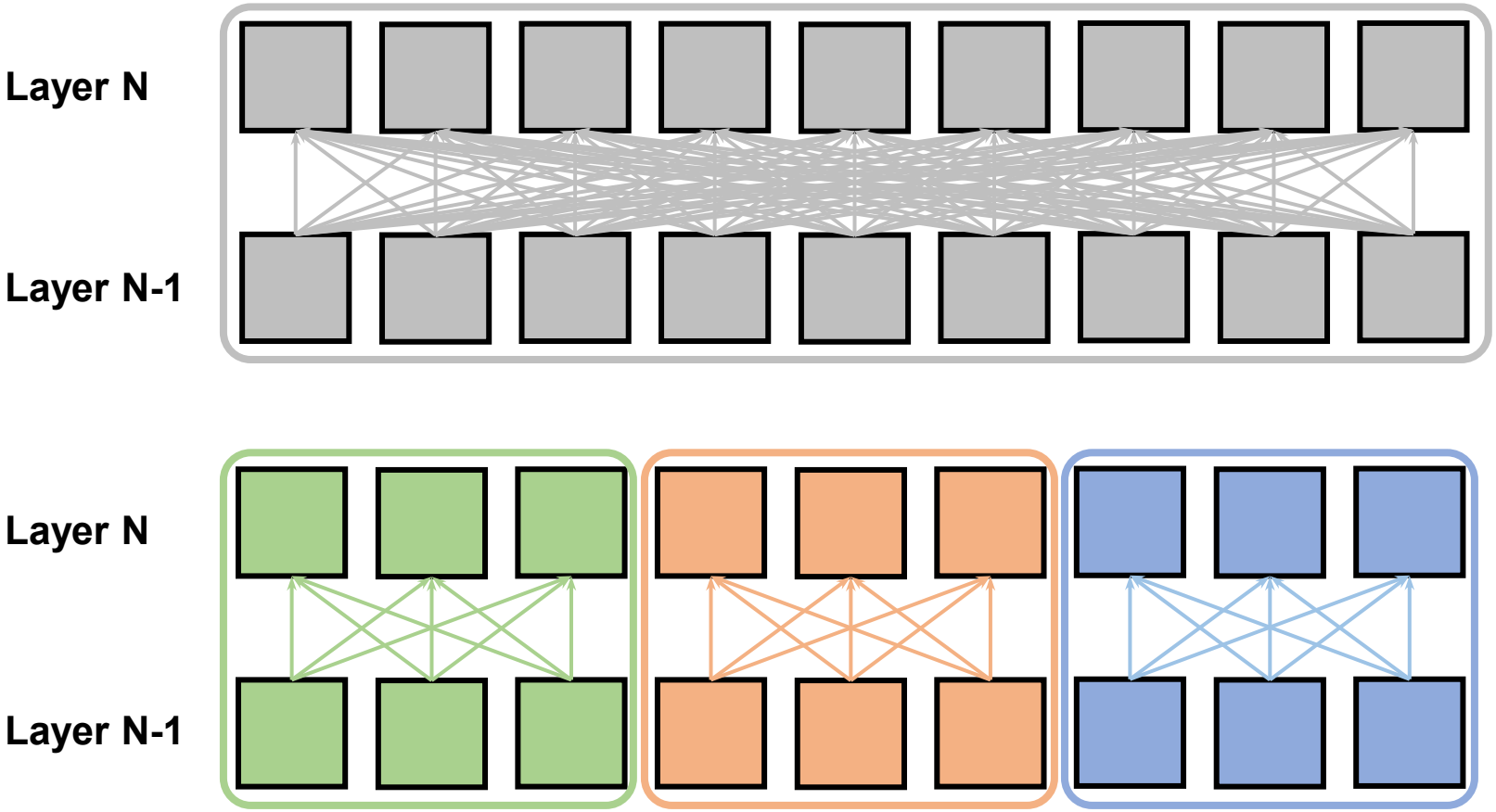
# Naïve Feature-masking is not Robust

- Example: Global attention in Vision Transformer




- Every token/feature receives signals from all tokens in the previous layer (global receptive field)
- **Large receptive fields can hurt robustness**
  - One localized corrupted token/feature can corrupt all tokens/features
  - **Corruption is no longer localized!** Masking no longer works :(

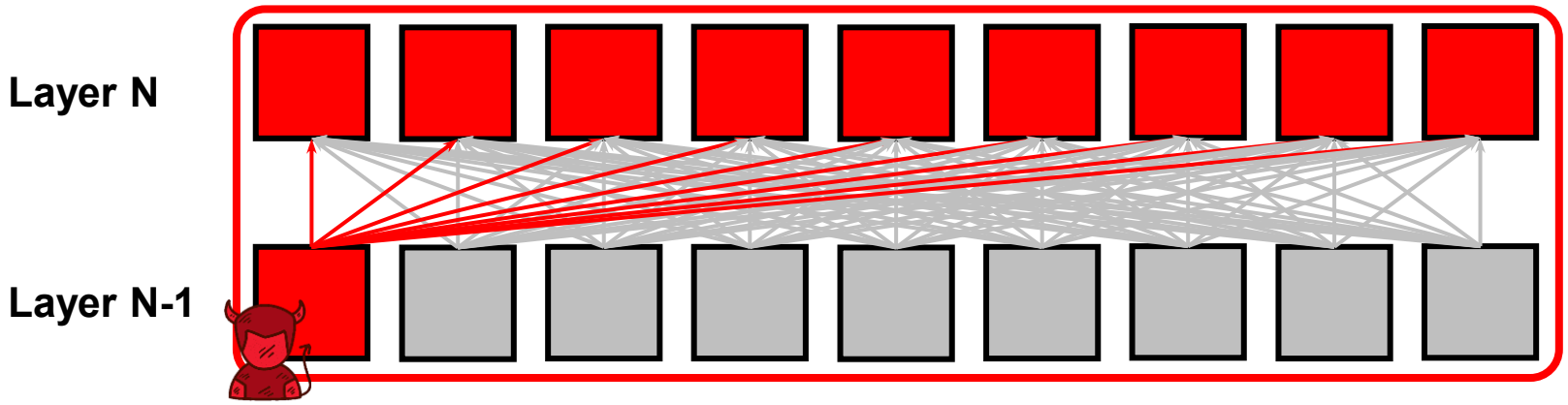
# Solution: Enforcing Small Receptive Fields before Masking



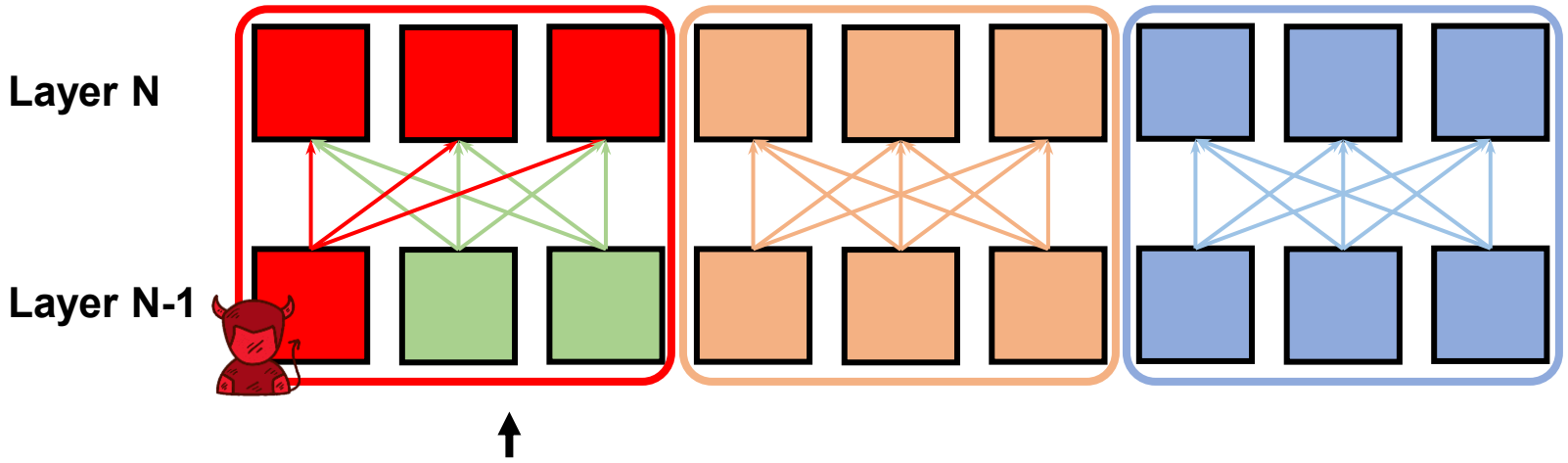

**Global Attention**: attention across all 9 visual tokens

**Local Attention**: attention within each sub-group of 3 visual tokens, but no inter-group attention

# Solution: Enforcing Small Receptive Fields before Masking



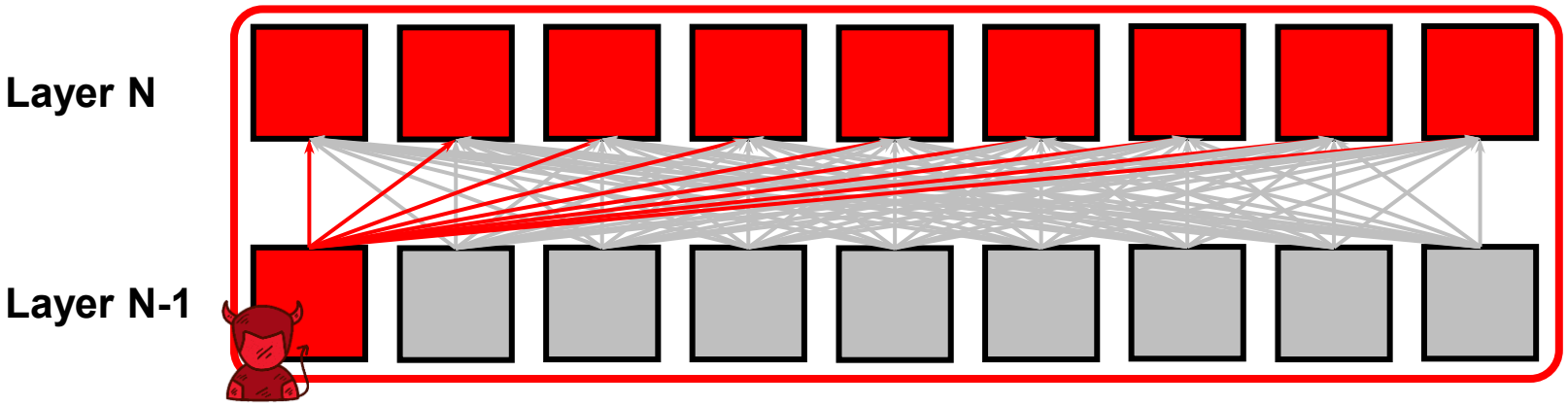

**Global Attention**: attention across all 9 visual tokens



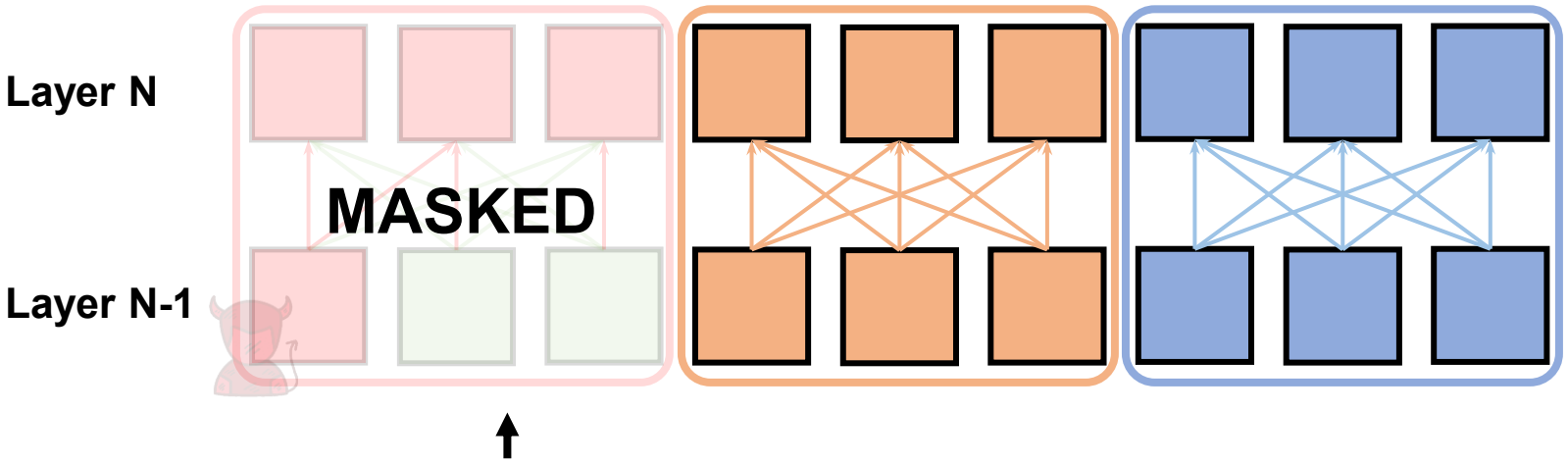

**Local Attention**: attention within each sub-group of 3 visual tokens, but no inter-group attention

**Masking can now provide robustness against this localized corruption!**

# Solution: Enforcing Small Receptive Fields before Masking




**Global Attention**: attention across all 9 visual tokens



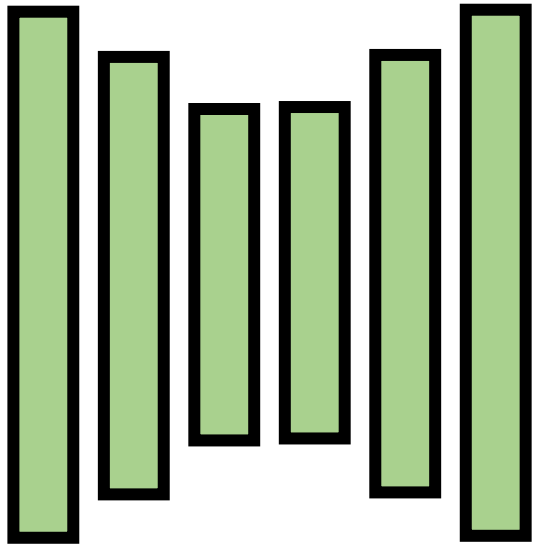

**Local Attention**: attention within each sub-group of 3 visual tokens, but no inter-group attention

**Masking can now provide robustness against this localized corruption!**

# PatchCURE: A Generalized Defense Framework

- Convert a vanilla undefended model into a defense model with tunable computation efficiency and certifiable robustness

Large Receptive Field (LRF) Layer

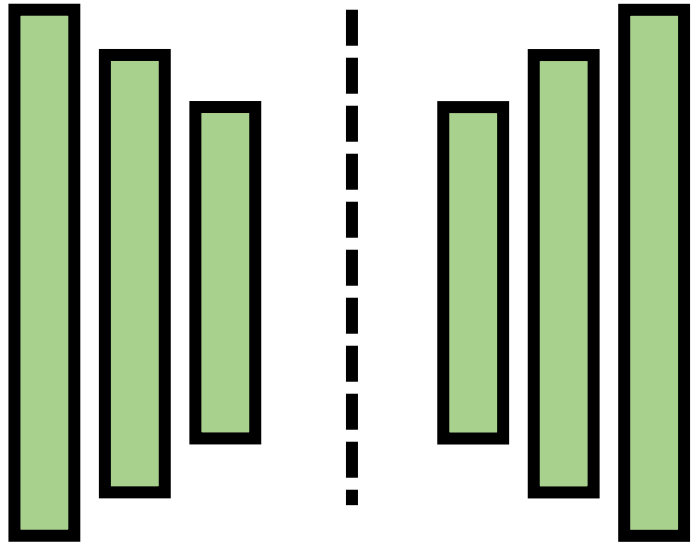


**Undefended model**

# PatchCURE: A Generalized Defense Framework

- Convert a vanilla undefended model into a defense model with tunable computation efficiency and certifiable robustness

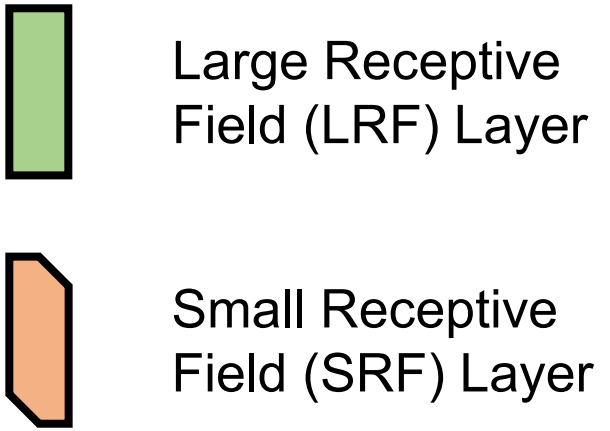
Large Receptive Field (LRF) Layer

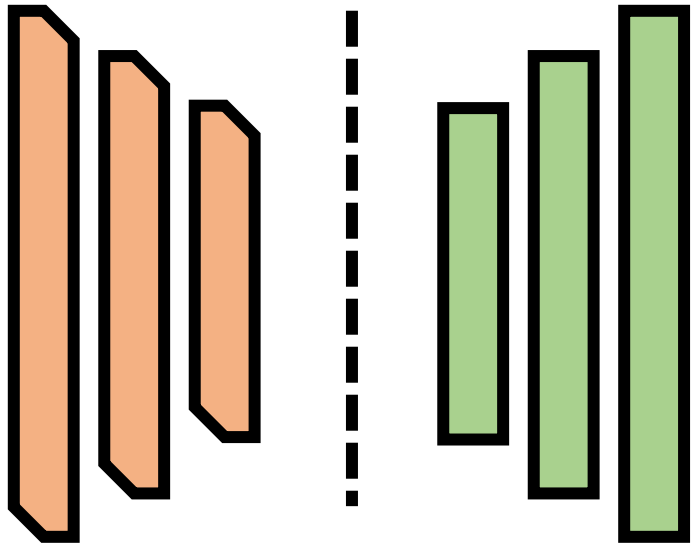


1. Split at $k^{th}$ layer

# PatchCURE: A Generalized Defense Framework

- Convert a vanilla undefended model into a defense model with tunable computation efficiency and certifiable robustness

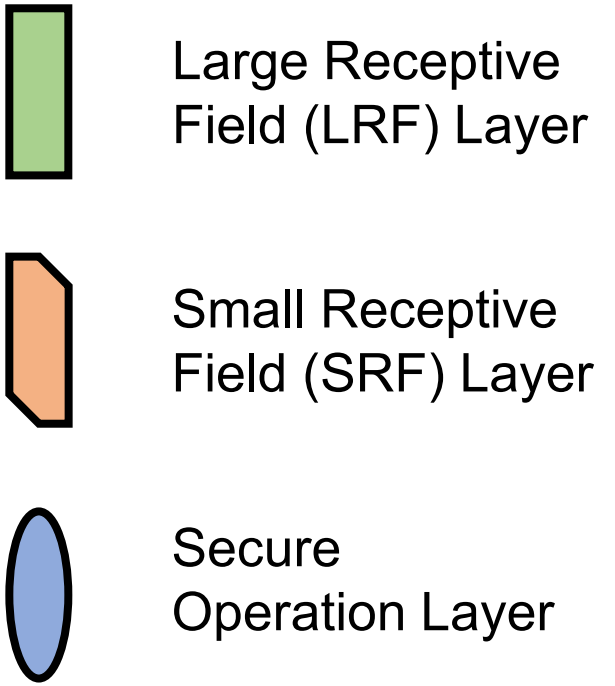


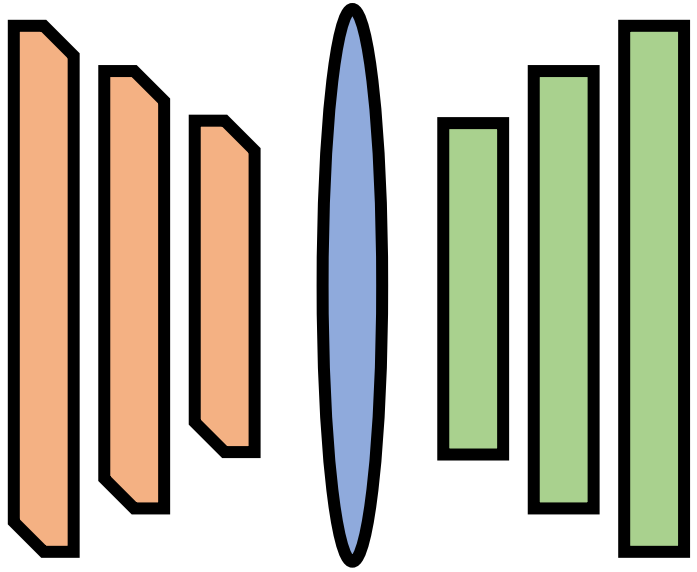



1. Split at $k^{th}$ layer
2. Convert LRF to SRF layers

# PatchCURE: A Generalized Defense Framework

- Convert a vanilla undefended model into a defense model with tunable computation efficiency and certifiable robustness



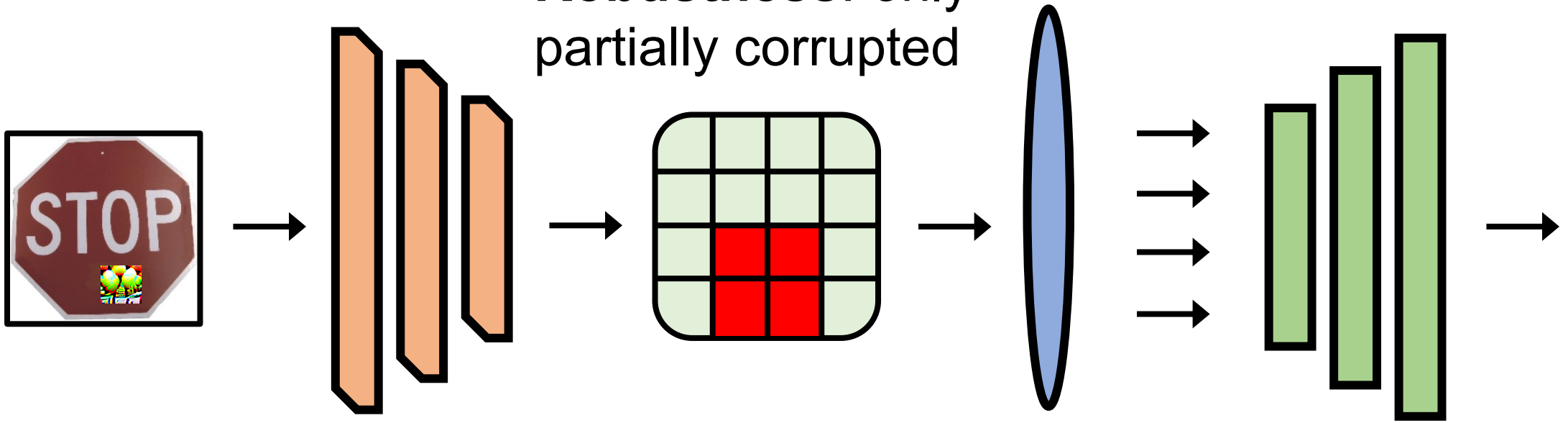



1. Split at $k^{th}$ layer
2. Convert LRF to SRF layers
3. Insert secure operation (e.g., masking-based analysis)

# PatchCURE Inference

**Efficiency**: multiple passes over only a small sub-model

**Robustness**: only partially corrupted




**Input Image**

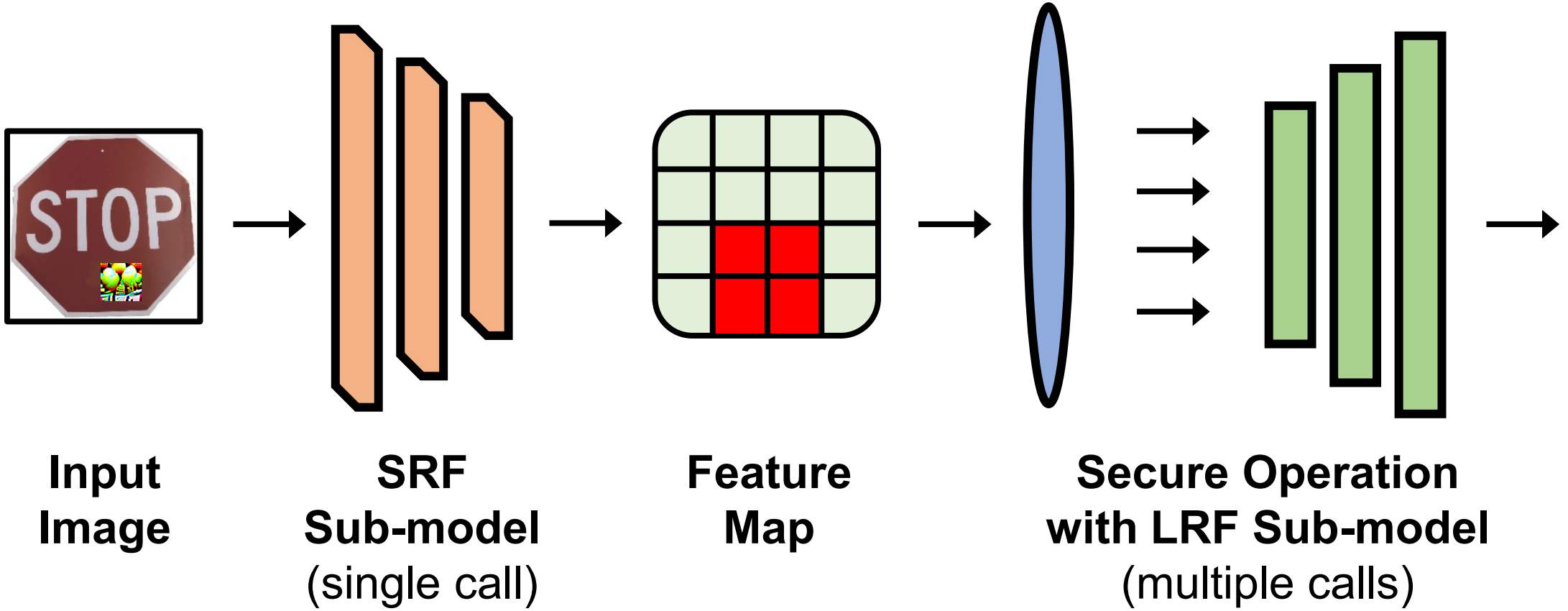
**SRF Sub-model** (single call)

**Feature Map**

**Secure Operation with LRF Sub-model** (multiple calls)

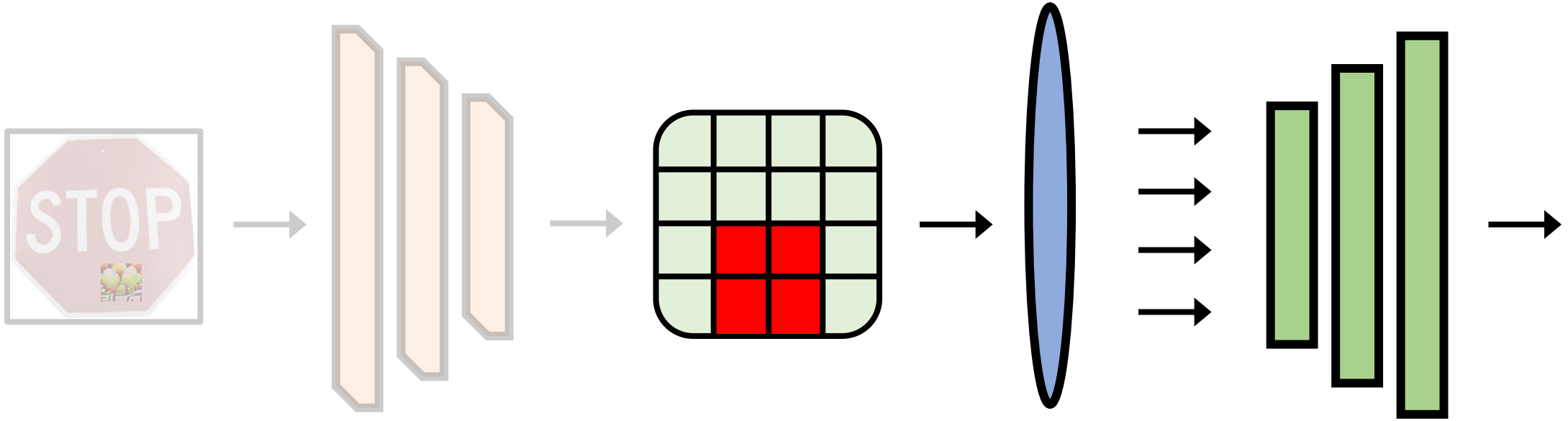
# PatchCURE Robustness Certification

- Treat feature map as the input image and directly apply off-the-shelf certification technique



# PatchCURE Robustness Certification

- Treat feature map as the input image and directly apply off-the-shelf certification technique



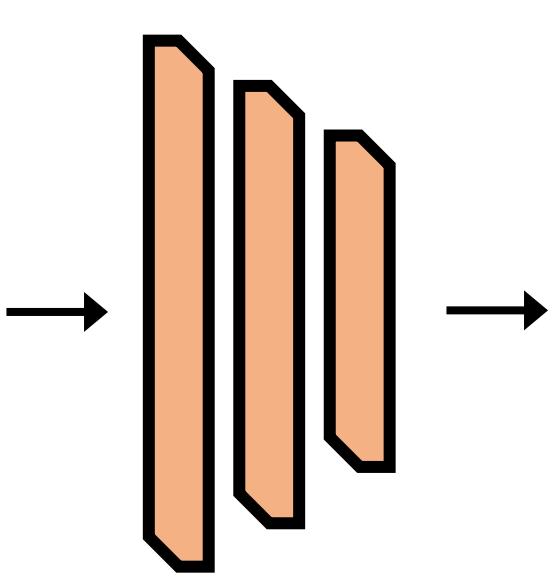

Input Image

SRF Sub-model (single call)

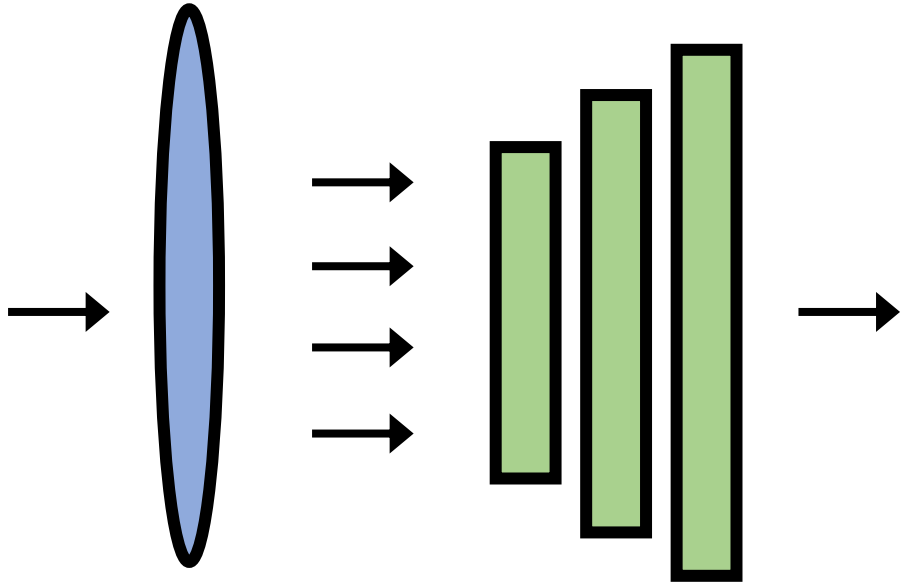
**Input "Image"**

**Secure Operation with LRF Sub-model** (multiple calls)

# PatchCURE Pipeline

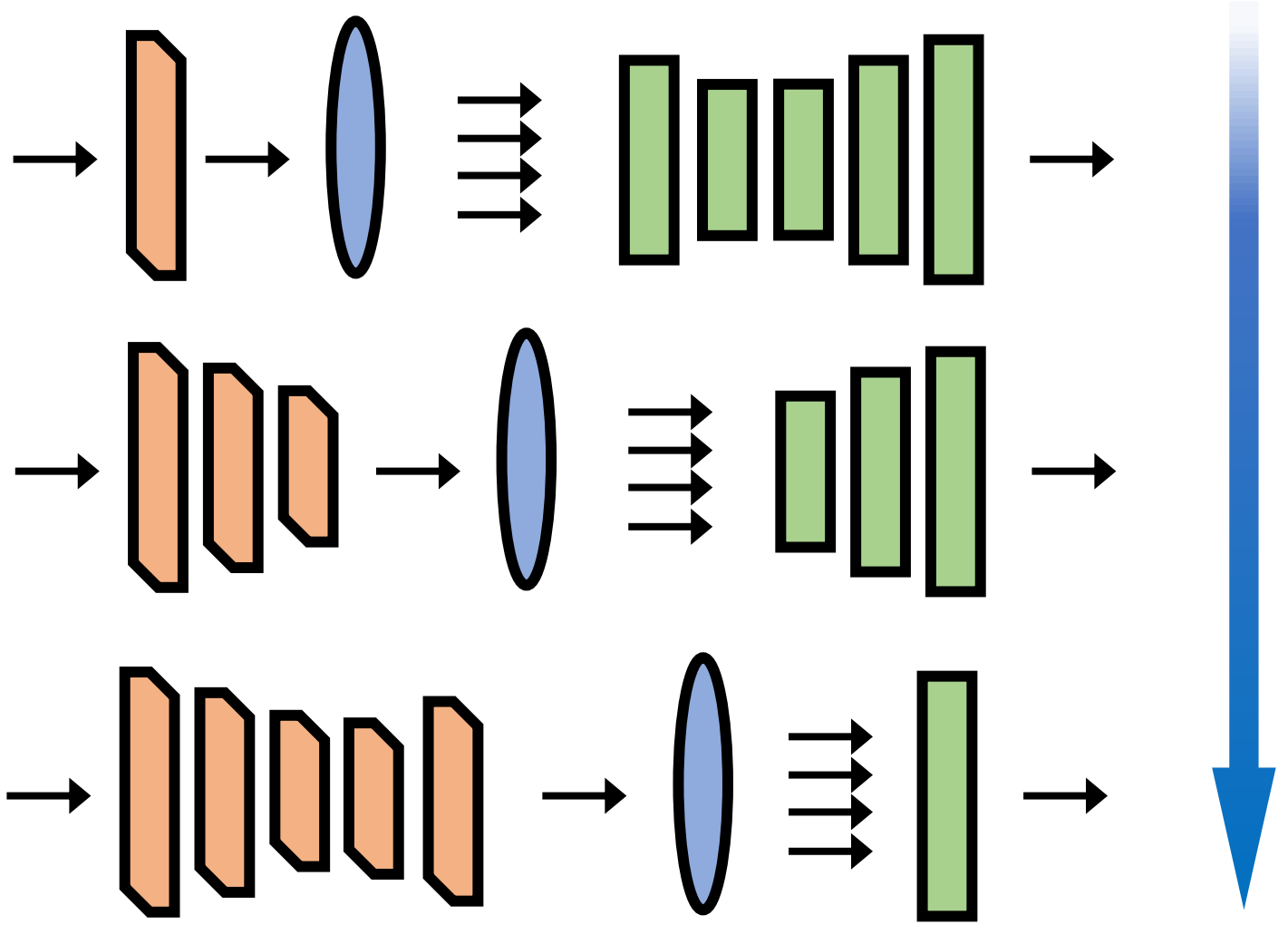


**SRF**
**Sub-model**
(single call)



**Secure Operation**
**with LRF Sub-model**
(multiple calls)

# Splitting Layer k Adjust the Defense Performance



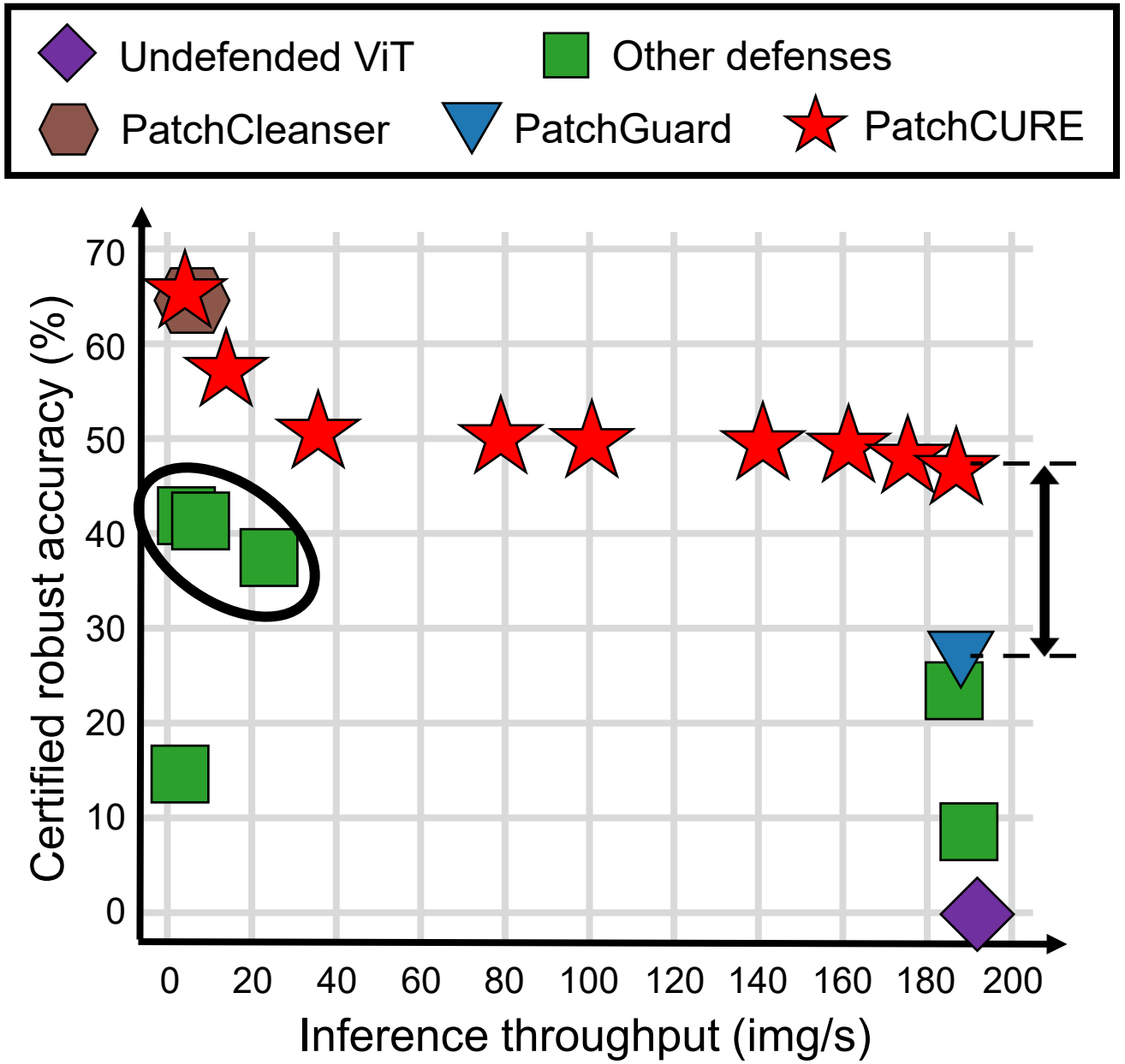
*Split at a deeper layer:*

**Computation efficiency** ⇧

**Certifiable robustness** ⇩

# ImageNet Evaluation

- Diverse robustness and efficiency
- Best across all efficiency levels
- Large robustness improvement (18%) for efficient defenses
- Efficient PatchCURE instances even outperform many inefficient prior works



# PatchCURE: An Extensible and Powerful Framework

| Defense | SRF sub-model | LRF sub-model | Secure operation | Splitting layer |
|---|---|---|---|---|
| **PatchCURE** | **ViT-SRF/BagNet** | **ViT/ResNet** | **Double-masking** | **Any layer** |
| PatchGuard | BagNet | Linear classifier | Robust masking | Last feature layer |
| PatchCleanser | None | Any model | Double-masking | Input layer |
| Clipped BagNet | BagNet | Linear classifier | Feature clipping | Last feature layer |
| Derandomized Smoothing | Pixel bands to ResNet | None | Majority voting | Output layer |
| PatchGuard++ | BagNet | Linear classifier | consistency check | Output layer |
| BagCert | Modified BagNet | None | Majority voting | Last layer |
| Randomized Cropping | Cropped images to ResNet | None | Majority voting | Last layer |
| ScaleCert | First few CNN layers | Remaining CNN layers | “SIN analysis” | First feature layer |
| Smoothed ViT | Pixel bands to ViT | None | Majority voting | Output layer |
| ECViT | Pixel bands to ViT | None | Majority voting | Output layer |
| ViP | Pixel bands to ViT | None | Majority voting | Output layer |
| Yatsura et al. | Pixel bands | None | Majority voting | Output layer |

(and more)

# PatchCURE Takeaways

- A **defense framework** with tunable certifiable robustness and computation efficiency
  - Feature-space defense with a combination of SRF and LRF techniques
  - State-of-the-art robustness across all efficiency levels
  - Subsume all existing defenses that are scalable to full-size ImageNet