CertMask: Certifiable Defense Against Adversarial Patches via Theoretically Optimal Mask Coverage

Xuntao Lyu[1*], Ching-Chi Lin[2*], Abdullah Al Arafat[1, 3], Georg von der Brüggen[2], Jian-Jia Chen[2, 4], Zhishan Guo[1, 2]

[1]North Carolina State University
[2]TU Dortmund University
[3]Florida International University
[4]Lamarr Institute for AI and ML
xlyu5@ncsu.edu, zguo32@ncsu.edu, chingchi.lin@tu-dortmund.de, georg.von-der-brueggen@tu-dortmund.de, jian-jia.chen@tu-dortmund.de, aarafat@fiu.edu

Abstract

Adversarial patch attacks inject localized perturbations into images to mislead deep vision models. These attacks can be physically deployed, posing serious risks to real-world applications. In this paper, we propose CertMask, a certifiably robust defense that constructs a provably sufficient set of binary masks to neutralize patch effects with strong theoretical guarantees. While the state-of-the-art approach (PatchCleanser) requires two rounds of masking and incurs $O(n^2)$ inference cost, CertMask performs only a single round of masking with $O(n)$ time complexity, where n is the cardinality of the mask set to cover an input image. Our proposed mask set is computed using a mathematically rigorous coverage strategy that ensures each possible patch location is covered at least k times, providing both efficiency and robustness. We offer a theoretical analysis of the coverage condition and prove its sufficiency for certification. Experiments on ImageNet, ImageNette, and CIFAR-10 show that CertMask improves certified robust accuracy by up to +13.4% over PatchCleanser, while maintaining clean accuracy nearly identical to the vanilla model.

Extended version — https://arxiv.org/abs/2511.09834

Introduction

Deep learning models have achieved strong performance in vision tasks such as robotics, autonomous driving, and surveillance. However, studies (Xiao et al. 2018; Yuan et al. 2022; Shi et al. 2022; Croce et al. 2022) show that these models are vulnerable to adversarial attacks. Among them, adversarial patch attacks (Brown et al. 2017; Yang et al. 2020; Karmon et al. 2018) are particularly concerning due to their spatial locality. Unlike traditional perturbations that modify the entire input, patch attacks introduce small, confined regions that can significantly degrade performance. Moreover, they pose a practical threat because they can be physically instantiated and deployed in real-world scenes (Yuan et al. 2024; Xiao et al. 2021; Wang et al. 2022).

Many existing defenses against adversarial patches require strong assumptions or are limited in practice. Some methods (Levine and Feizi 2020; Lin et al. 2021; Metzen and Yatsura 2021; Xiang et al. 2021) rely on extracting intermediate activations from the model and enforcing architectural constraints, such as limiting the receptive field to reduce sensitivity to localized perturbations. These approaches typically assume access to the internal structure of the model and are restricted to specific architectures. Furthermore, constraining the receptive field can lead to a decline in clean accuracy, and the high computational cost of these methods often limits their applicability to low-resolution datasets or controlled offline settings.

Certifiably robust defenses aim to overcome these limitations. They (Xiang, Mahloujifar, and Mittal 2022; Saha et al. 2023; Li, Zhang, and Xie 2022) aim to provide provable guarantees that the model's prediction will remain unchanged under worst-case patch attacks. PatchCleanser (Xiang, Mahloujifar, and Mittal 2022), one of the most effective examples, applies two rounds of masking to the input and aggregates predictions to certify robustness, is model-agnostic, and does not require retraining. However, PatchCleanser and similar defenses suffer from substantial computational overhead. In particular, PatchCleanser performs $O(n^2)$ forward passes (n is the number of masks to fully cover an input image) due to its pairwise masking strategy, making it impractical for high-resolution inputs or real-time deployment, where patch attacks can be executed instantly.

To address these limitations, we introduce CertMask, a novel certifiably robust defense that preserves provable robustness guarantees while achieving substantially improved computational efficiency. CertMask (illustrated in Figure 1) deterministically allocates the provably sufficient number of masks required to ensure that any adversarial patch of bounded size is *fully covered at least k times* (a.k.a *k-fold coverage*). By abstracting the patch coverage requirement into a discrete dot coverage problem, we derive the provably sufficient number of masks needed to ensure k-fold coverage for all potential patch locations. Both necessary and sufficient conditions for certified robustness are estab-

*These authors contributed equally.

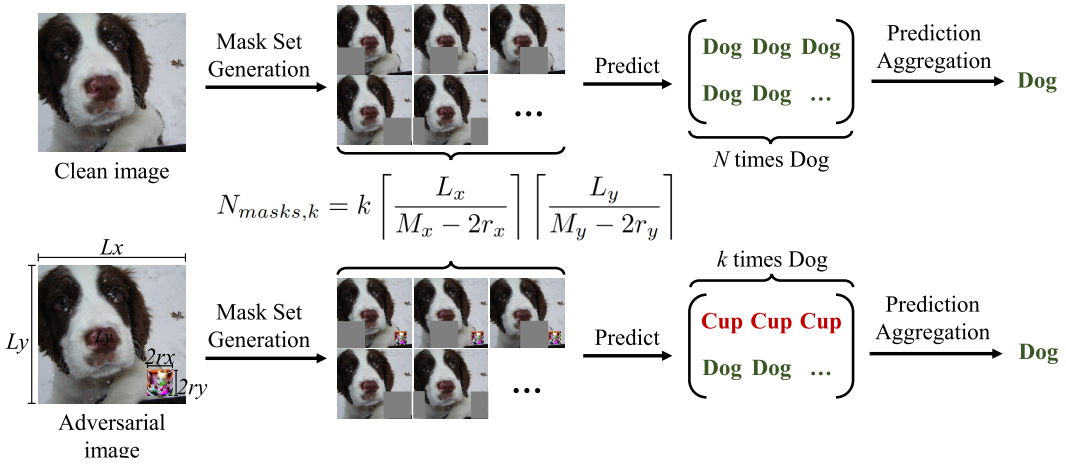

Figure 1: Overview of the CertMask inference pipeline. Given an input image of size $L_x \times L_y$, we deterministically construct a set of $N_{masks,k}$ binary masks, where each mask has spatial support $M_x \times M_y$ and is positioned such that every patch of size at most $2r_x \times 2r_y$ is guaranteed to be covered by exactly k different masks. Each masked image is evaluated by the classifier, yielding a prediction. For aggregation, if all predictions agree, we output the unanimous result. If disagreement occurs and one class appears exactly k times, that class is returned as the certified prediction. Otherwise, the majority class is selected to account for potential benign misclassifications.

lished. Based on those conditions, we form two constructions with asymptotically optimal inference complexity of $O(n)$, a phenomenal improvement over prior $O(n^2)$ methods.

Model and Problem

This section introduces the system model, defines the threat model, and formally states the problem to be solved.

System Model

The **target domain** is a 2-dimensional spatial region where an adversarial patch may appear, such as an input image. This domain is inherently **discrete**, composed of $L_x \times L_y$ pixels, forming a continuous rectangular area from $[0, L_x]$ and $[0, L_y]$ along the x- and y-axis, respectively.

An **adversarial patch** is a localized region within the target domain, designed to mislead a deep learning model. We model this patch as a solid, rectangular area with known dimensions but an **unknown location** within the target domain. Let the patch be centered at (C_x, C_y) and have a total width of $2r_x$ and a total height of $2r_y$ (i.e., extending from $C_x - r_x$ to $C_x + r_x$ and from $C_y - r_y$ to $C_y + r_y$).

A **mask** is a defensive mechanism applied to the target domain, intended to neutralize or obscure potential adversarial patches. In our model, a mask is a rectangular region of fixed dimensions, $M_x \times M_y$. The center of a mask can be positioned anywhere in the target domain (hence, its boundaries may partially extend beyond the defined target domain). When applied, a mask effectively leads to ignorance of the information within its boundaries, preventing it from influencing the deep learning model's output.

Threat Model

We consider a test-time adversarial patch threat model. Let f be a classifier and $\mathbf{x} \in \mathbf{R}^{L_x \times L_y \times C}$ (where L_x and L_y denote the image width and height, and C is the number of channels) be an input image with ground-truth label y. An adversary aims to construct an adversarial example $\mathbf{x}'$ such that $f(\mathbf{x}') \neq y$, by modifying a localized spatial region of $\mathbf{x}$.

The adversarial modification is constrained to a rectangular patch region $\Omega \subset [0, L_x] \times [0, L_y]$ with known size $2r_x \times 2r_y$, placed arbitrarily within the image. This constraint is represented by a binary mask $\mathbf{r} \in \{0,1\}^{L_x \times L_y \times C}$, where $\mathbf{r}_{i,j} = 1$ if $(i, j) \in \Omega$, and 0 otherwise. The adversarial example is defined as:

$$\mathbf{x}' = \mathbf{r} \odot \mathbf{z} + (1 - \mathbf{r}) \odot \mathbf{x}, \tag{1}$$

where $\mathbf{z} \in \mathbf{R}^{L_x \times L_y \times C}$ denotes arbitrary adversarial content, and $\odot$ is element-wise multiplication.

We make no assumptions about the attacker's generation method or patch location. We only assume a known patch size, specifically that the adversarial region has dimensions $2r_x \times 2r_y$. This allows CertMask to defend against highly adaptive and unrestricted patch attacks without requiring access to patch placement or construction details.

Adversarial Patch Covering (APC) Problem

We define the Adversarial Patch Covering **(APC)** problem as follows: Consider a 2-dimensional spatial domain, representing an image or sensor input, which may contain a single adversarial patch of unknown location. The objective is to determine the minimum number of identical rectangular masks that must be strategically applied to this domain to guarantee that the patch is fully covered by at least k of these masks. An adversarial patch is **fully covered** by a mask if

and only if the entire area of the patch is strictly contained within the boundaries of the mask (so it can be neutralized).

Definition 1 (Fully Covered): A mask spanning from $[x, x+M_x]$ and $[y, y+M_y]$ fully covers an adversarial patch centered at (C_x, C_y) with radii r_x, r_y if:

$$x \leq C_x - r_x \text{ and } C_x + r_x \leq x + M_x, \text{ and}$$
$$y \leq C_y - r_y \text{ and } C_y + r_y \leq y + M_y$$

For analytical tractability, we assume that all masks are of identical, rectangular dimensions, that the masks are sufficiently large to cover the adversarial patch, and that the adversarial patch itself is also rectangular.

This problem directly relates to practical challenges in ensuring the robustness of deep learning models against adversarial patch attacks. By finding the minimum number of masks needed to cover an unknown adversarial patch at least k times, we are essentially determining an efficient masking strategy to neutralize its effect, thus preventing the attack from influencing the model's output without incurring excessive computational cost.

Covering an Adversarial Patch

We start by examining patch coverage in a 1-dimensional domain, then extend this analysis to 2-dimensional scenarios. We derive the conditions required for a patch to be fully covered by a mask—they are fundamental building blocks for developing effective solutions to the APC problem.

APC in the 1-Dimensional Domain

We begin by considering patch covering in a discrete 1-dimensional domain. Assume an adversarial patch with radius r centered at C, spanning the interval $[C-r, C+r]$. A mask of size $M > 2r$ is used for covering. To fully cover this patch, the mask's left endpoint, x, must satisfy $x \leq C - r$, while its right endpoint, $x+M$, must satisfy $x+M \geq C+r$.

As shown below, this patch-covering problem can be simplified into a dot-covering problem, where the full coverage of a patch is determined by the position of its center C.

Theorem 1. An adversarial patch centered at C is fully covered by a mask $[x, x+M]$ if and only if $x + r \leq C \leq x + M - r$.

Proof. Provided in Appendix A in the extended version.

Based on Theorem 1, we can now define an *effective coverage* interval within a mask. If the center C of an adversarial patch falls within this interval, the patch is guaranteed to be fully covered by the mask, as shown in Figure 2.

APC in the 2-Dimensional Domain

As established in Definition 1, an adversarial patch is **fully covered** by a mask if and only if its entire area is strictly contained within the mask's boundaries. This requires that the conditions for the patch's extent relative to the mask's boundaries are met independently along both the x- and y-axes in a 2-dimensional domain.

Similar to the 1-D case, 2-D APC can be simplified by focusing on the center (C_x, C_y) of the adversarial patch. Since the coverage in each dimension is independent, by Theorem 1, $x + r_x \leq C_x \leq x + M_x - r_x$ and $y + r_y \leq C_y \leq y + M_y - r_y$ are conditions for 2-D coverage. These naturally lead to a rectangular *effective coverage* area within a mask: if the center (C_x, C_y) of the adversarial patch falls within this area, the entire patch is fully covered by the mask (see Figure 3).

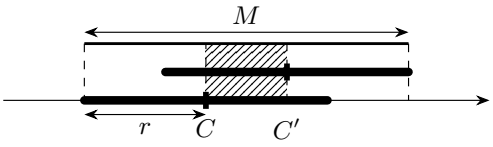

Figure 2: **1-D Effective Coverage Interval**. For a mask (size M) and adversarial patch (radius r), the effective coverage interval (shaded) has a size of $M - 2r$, representing the range where full patch coverage is guaranteed.

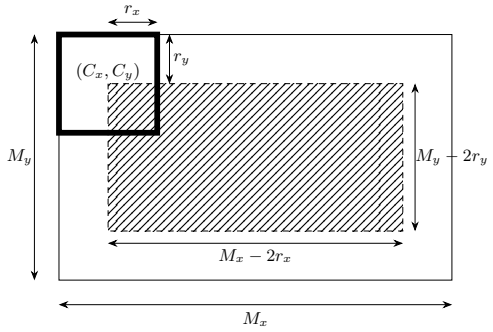

Figure 3: **2-D Effective Coverage Area**. The shaded region illustrates the effective coverage area for a mask of size $M_x \times M_y$. For an adversarial patch with radii r_x and r_y, this area has dimensions $(M_x - 2r_x) \times (M_y - 2r_y)$.

Methodology: CertMask

This section presents our methodology for addressing the *APC Problem*, which aims to minimize the number of masks applied while ensuring that any adversarial patch is fully covered by at least k distinct masks, regardless of its unknown location. We first analyze the basic case of single coverage ($k = 1$), and then extend our formulation to the general k-coverage setting.

Single Cover

Single Cover in the 1-D domain. Theorem 1 shows that an adversarial patch of radius r is fully covered by a mask of size M if and only if the patch's center C falls within the mask's effective coverage interval $[x + r, x + M - r]$. This effective coverage interval has a length of $M - 2r$. Based on this crucial observation, we derive the following optimal pavement strategy for a 1-dimensional domain $[0, L]$.

Our strategy involves tiling the target domain with these effective coverage intervals, placing them contiguously without any gaps in the effective coverage as shown in Figure 4. While the effective coverage intervals do not overlap,

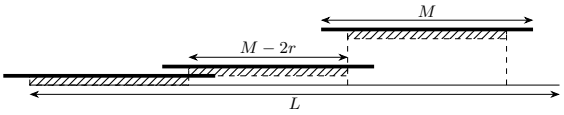

Figure 4: **Pavement strategy for 1-D domain**. This figure illustrates the mask placement strategy where effective coverage intervals are arranged contiguously, ensuring complete and gap-free coverage of the domain.

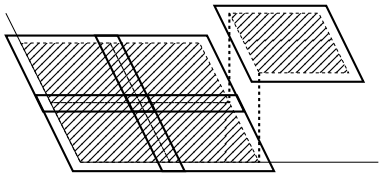

Figure 5: **Pavement strategy for 2-D domain**. This figure illustrates the mask placement strategy where effective coverage areas are arranged contiguously, ensuring complete and gap-free coverage of the 2-D domain.

the masks themselves will inherently overlap by a length of $2r$ at their boundaries due to the r-offset from the mask edge to the effective coverage edge.

To fully cover the target domain $[0, L]$ with these effective coverage intervals, each of length $M - 2r$, the minimum number of masks required is $\lceil \frac{L}{M-2r} \rceil$. We now formally prove that this quantity represents the optimal (minimum) number of masks for the 1-D APC with $k = 1$.

Theorem 2. To ensure that every possible position of an adversarial patch with length $2r$ within a 1-D domain of length L is fully covered by at least one mask of size M, the number of masks N_{masks} must satisfy:

$$N_{masks} \geq \left\lceil \frac{L}{M - 2r} \right\rceil \tag{2}$$

Proof. Provided in Appendix A in the extended version.

Single Cover in the 2-D domain. We extend the optimal pavement strategy to a 2-dimensional domain based on our 1-D findings. As established in our preliminary analysis, a patch is fully covered if and only if its center (C_x, C_y) falls into the effective coverage area of one of the masks. For a 2-D mask, this effective coverage area is a rectangle of size $(M_x - 2r_x)(M_y - 2r_y)$.

An adversarial patch is covered by at least one mask on a 2-dimensional rectangular domain if and only if the union of the effective coverage areas of the deployed masks completely covers the entire $L_x \times L_y$ domain. Since each effective coverage area has dimensions $M_x - 2r_x$ by $M_y - 2r_y$, the total area to be covered is $L_x L_y$. Therefore, the theoretical minimum number of masks, N^{opt}_{masks}, must satisfy:

$$N^{opt}_{masks} \geq \left\lceil \frac{L_x L_y}{(M_x - 2r_x)(M_y - 2r_y)} \right\rceil \tag{3}$$

The pavement strategy for 2-D is a direct extension of the 1-D approach. We tile the 2-D domain by arranging masks such that their effective coverage areas are contiguously placed. Starting from the lower-left corner of the target domain, masks are applied row by row and column by column. Along the x-axis, masks are placed such that their effective coverage areas are adjacent, continuing until the total covered length exceeds L_x. The same principle applies along the y-axis. This strategy requires $\lceil \frac{L_x}{M_x - 2r_x} \rceil$ masks along the x-dimension and $\lceil \frac{L_y}{M_y - 2r_y} \rceil$ masks along the y-dimension. The total number of masks used by this strategy, $\lceil \frac{L_x}{M_x - 2r_x} \rceil \lceil \frac{L_y}{M_y - 2r_y} \rceil$, closely approximates the derived minimal bound.

Multiple Cover

Having established the mask deployment strategies for single-cover scenarios, we now extend our analysis to cases where an adversarial patch must be fully covered by more than one mask, specifically by at least k masks ($k > 1$). Empirical evidence (McCoyd et al. 2020; Xiang, Mahloujifar, and Mittal 2022) suggests that increasing this coverage multiplicity k can significantly enhance the robustness of deep learning models against adversarial patch attacks.

To achieve the desired k-fold coverage for an adversarial patch, we propose two distinct pavement strategies. **Replicated Tiling** is a straightforward extension of our single-cover solution and establishes a fundamental baseline. **Offset Tiling** provides an alternative and more intricate approach to mask placement, exploring more nuanced optimizations for practical deployment.

Strategy 1: Replicated Tiling. From our analysis of 2-D single coverage, we established a 2-approximation tiling strategy that fully covers each adversarial patch at least once. A direct and intuitive approach to achieve k-fold coverage is to simply replicate this single-cover pavement plan k times. This ensures that any adversarial patch, regardless of its location, will be fully covered by k identical masks.

We first establish a fundamental theoretical lower bound for the minimum number of masks required, $N^{opt}_{masks,k}$. Subsequently, we formally prove the approximation ratio of the **Replicated Tiling** strategy relative to this lower bound.

Theorem 3. To ensure that every possible position of a rectangular adversarial patch with size $2r_x \times 2r_y$ within a 2-D spatial domain of size $L_x \times L_y$ is fully covered by at least k masks of size $M_x \times M_y$, the minimum number of masks $N^{opt}_{masks,k}$ must satisfy:

$$N^{opt}_{masks,k} \geq k \left\lceil \frac{L_x L_y}{(M_x - 2r_x)(M_y - 2r_y)} \right\rceil \tag{4}$$

Proof. Provided in Appendix A in the extended version.

Theorem 4. The **Replicated Tiling** strategy provides an approximation ratio of 2.

Proof. Provided in Appendix A in the extended version.

In practical implementation, instead of physically duplicating masks, this can correspond to applying the same masking pattern iteratively over k different processing rounds or frames, or using k independent masking layers.

Strategy 2: Offset Tiling. We propose a compact and evenly distributed mask placement strategy for k-fold coverage over a 2-D spatial domain termed Offset Tiling. Unlike Replicated Tiling, Offset Tiling interleaves multiple shifted mask grids within the same spatial domain to ensure that every adversarial patch is fully covered by at least k different masks.

To achieve k-fold coverage, we decompose k into two positive integers: $k = m \cdot n,$ where, $m, n \in \mathbb{Z}^+$. Here, m and n are the desired number of overlapping masks along the horizontal and vertical axes. We then arrange the effective coverage regions of the masks on a uniform grid with horizontal and vertical strides:

$$s_x = \frac{M_x - 2r_x}{m}; \quad s_y = \frac{M_y - 2r_y}{n}.$$

We construct the tiling by placing masks such that the centers of their effective coverage areas are positioned at: $(x_i, y_j) = (i \cdot s_x, j \cdot s_y), \quad$ for integers $i, j \geq 0$. This guarantees that each point in the image will be guaranteed to be within the effective coverage area of at least m horizontally placed masks and n vertically placed masks, achieving the required k-fold coverage in total.

To ensure full coverage near the image boundaries, we adopt a wrap-around strategy: when a mask extends beyond the image border, its overflow wraps around to the opposite side (i.e., toroidal padding). This avoids coverage gaps at the edges and guarantees uniformity across the domain.

Total number of masks required under Offset Tiling is:

$$N_{\text{masks},k} \geq \left\lceil \frac{m \cdot L_x}{M_x - 2r_x} \right\rceil \cdot \left\lceil \frac{n \cdot L_y}{M_y - 2r_y} \right\rceil \tag{5}$$

This approach provides a compact and efficient method for k-fold certified coverage, distributing masks more evenly than simple replication.

Experiments

Setup

In this section, we describe the evaluation setup including the datasets, models, attack configurations, evaluation metrics, and baseline defenses used to assess the effectiveness and robustness of our method.

Datasets and Models. Three image classification benchmarks are used: ImageNet (Deng et al. 2009), ImageNette (fast.ai 2020), and CIFAR-10 (Krizhevsky, Hinton et al. 2009). We use three image classification architectures: ResNetV2-50 (He et al. 2016), ViT-B/16 (Dosovitskiy et al. 2020), and ResMLP-S24 (Touvron et al. 2022). All models are pretrained on ImageNet and fine-tuned with Cutout augmentation as in PatchCleanserw (Xiang, Mahloujifar, and Mittal 2022). We resize images to 224×224 via bicubic interpolation.

Patch Attacks. Based on patch sizes commonly adopted in prior works (Chiang et al. 2020; Xiang, Mahloujifar, and Mittal 2022), we evaluate certified robustness against square patches occupying 1%, 2%, and 3% of input pixels for ImageNet and ImageNette, and 0.4% and 2.4% for CIFAR-10. Patches are allowed at arbitrary locations with unrestricted content within image bounds.

Metrics. We report (i) **clean accuracy**, which is the fraction of unperturbed test images classified correctly, and (ii) **certified robust accuracy**, which is the fraction of images where our certification procedure verifies prediction invariance against any patch within the threat model.

Baselines. We compare with the state-of-the-art method PatchCleanser (PC) (Xiang, Mahloujifar, and Mittal 2022) and other certified defenses, namely PatchGuard (PG) (Xiang et al. 2021), Distribution Smoothing (DS) (Levine and Feizi 2020), and Clipped BagNet (CBN) (Zhang et al. 2020), using their best reported configurations.

Results

Clean and certified robust accuracy. Table **??** reports the clean accuracy and certified robust accuracy of CertMask (CM), compared against the baseline approaches. All methods are evaluated under a unified protocol, using the same models, datasets, and patch sizes. For each patch threat level, we adopt a square mask with side length 16, 32, 48, or 56 to correspond to 0.4%, 1%, 2%, and $\{2.4\%, 3\%\}$ pixel patches, respectively. A fixed $3 \times 2 = 6$-fold mask coverage is implemented using our Offset Tiling strategy. For fair comparison, we use the same attack settings across methods, and highlight in bold the best certified accuracy achieved by both CertMask and the baseline.

CertMask consistently achieves competitive or superior clean accuracy across all datasets and architectures. As shown in Table **??**, CM-ViT attains 99.6% clean accuracy on ImageNette and 99.0% on CIFAR-10, matching PC-ViT. These values are also close to the clean performance of the corresponding undefended models reported in Table **??**, indicating that CertMask introduces minimal accuracy degradation. For example, CM-ViT maintains only a 0.2% drop from the vanilla 99.8% on ImageNette and exactly matches the baseline performance on CIFAR-10.

Certified robustness with minimal occlusion. In terms of certified robust accuracy, CertMask achieves substantial improvements over all prior methods. On ImageNet with a 2%-pixel adversarial patch, CM-ViT achieves 75.5% certified accuracy, surpassing PC-ViT (62.1%) by +13.4% and PG-DS (15.7%) by +59.8%. Similarly, on CIFAR-10 under a 2.4%-pixel patch, CM-ViT achieves 96.4%, outperforming PC-ViT (89.1%) by +7.3%. Similar gains are observed for ResNet and MLP across all datasets and threat models, reflecting the high efficacy of our mask construction.

A key factor for these gains is CertMask's single-round deterministic tiling strategy. Unlike PatchCleanser, which applies two rounds of randomized masking and requires $O(n^2)$ forward passes, CertMask ensures exact k-fold patch coverage in just one deterministic pass, using only $O(n)$ masks. This significantly reduces the total occlusion area, preserving more semantic context for prediction and improving robustness against patch attacks. Furthermore, the deterministic design eliminates the randomness-induced certification failures observed in prior stochastic methods.

Cross-architecture and cross-dataset generality. CertMask demonstrates strong generalization across architectures and datasets. On ImageNette with a 2%-pixel patch,

Method	ImageNette						ImageNet						CIFAR-10			
	1%		2%		3%		1%		2%		3%		0.4%		2.4%	
	clean	robust	clean	robust	clean	robust	clean	robust	clean	robust	clean	robust	clean	robust	clean	robust
CM-ResNet	**99.8**	**99.2**	**99.7**	98.9	**99.7**	98.6	81.9	72.2	81.7	69.8	81.7	68.5	97.9	94.6	98.0	92.4
CM-ViT	99.6	99.1	99.6	**99.0**	99.6	**98.9**	**84.1**	**77.6**	**84.5**	**75.5**	**84.5**	**74.0**	**99.0**	**97.7**	**98.8**	**96.4**
CM-MLP	99.4	98.9	99.4	98.7	99.4	98.5	80.2	72.1	80.0	69.8	79.8	67.9	97.5	93.8	97.1	90.8
PC-ResNet	**99.6**	96.4	**99.6**	94.4	**99.5**	93.5	81.7	58.4	81.6	53.0	81.4	50.0	98.0	88.5	97.8	78.8
PC-ViT	**99.6**	**97.5**	**99.6**	**96.4**	**99.5**	**95.3**	**84.1**	**66.4**	**83.9**	**62.1**	**83.8**	**59.0**	**99.0**	**94.3**	**98.7**	**89.1**
PC-MLP	99.4	96.8	99.3	95.1	99.4	94.6	79.6	58.4	79.4	53.8	79.3	50.7	97.4	86.1	97.0	78.0
CBN	94.9	74.6	94.9	69.9	94.9	45.9	49.5	13.4	49.5	7.1	49.5	3.1	84.2	44.2	84.2	9.3
DS	92.1	82.3	92.1	79.1	92.1	75.7	44.4	17.7	44.4	14.0	44.4	11.2	83.9	68.9	83.9	56.2
PG-BN	95.2	89.0	95.0	86.7	94.8	83.0	55.1	32.3	54.6	26.0	54.1	19.7	84.5	63.8	83.9	47.3
PG-DS	92.3	83.1	92.1	79.9	92.1	76.8	44.1	19.7	43.6	15.7	43.0	12.5	84.7	69.2	84.6	57.7

Table 1: Clean accuracy and certified robust accuracy (%) under $k = 6$ mask coverage ($k = 3 \times 2$). We use fixed square masks with side length 16 for 0.4% patch, 32 for 1% patch, 48 for 2% patch, and 56 for both 2.4% and 3% patch. Bold highlights the best result among CertMask (CM) and prior defenses.

	ImageNette	ImageNet	CIFAR-10
ResNet	99.8%	82.3%	98.3%
ViT	99.8%	84.8%	99.0%
MLP	99.5%	80.2%	97.8%

Table 2: Clean accuracy of vanilla models.

Model	Tiling	ImageNette		ImageNet		CIFAR-10	
		clean	robust	clean	robust	clean	robust
ResNet	Replicated	99.5	**99.3**	77.9	**73.7**	96.1	**94.9**
	Offset	**99.7**	98.6	**81.7**	68.5	**98.0**	92.4
ViT	Replicated	99.4	**99.2**	82.0	**78.6**	98.4	**97.8**
	Offset	**99.6**	98.9	**84.5**	74.0	**98.8**	96.4
MLP	Replicated	99.2	**99.0**	76.7	**72.5**	95.7	**94.1**
	Offset	**99.4**	98.5	**79.8**	67.9	**97.1**	90.8

Table 3: Effect of Tiling Strategy on Performance.

CertMask improves ViT's certified accuracy from 96.4% (PC-ViT) to 99.0%, and boosts ResNet's robustness from 94.4% to 98.9%. On ImageNet, CM-ViT improves certified robustness by over 10% absolute margin across all tested patch sizes. These improvements are consistent across CIFAR-10 as well, showing CertMask's adaptability to both low- and high-resolution images under varying threat levels.

Overall, CertMask sets a new state-of-the-art for certified patch robustness. Its deterministic, geometry-aware mask deployment achieves strong robust accuracy with minimal clean accuracy loss, offering a practical and certifiable solution for patch-resilient vision systems.

Detailed Parameter Analysis

In this subsection, we systematically analyze how core design choices—tiling strategy, coverage multiplicity (k), mask size, and patch size—affect the clean and certified robust accuracy of CertMask.

k	1	2	3	4	5	6
Clean (%)	82.0	80.8	82.5	83.4	83.8	84.5
Robust (%)	78.6	77.0	76.2	74.9	75.7	74.0

Table 4: Clean and robust accuracy under varying k.

Analysis of the performance of two Tiling Strategies. Table **??** compares Replicated Tiling and Offset Tiling under identical evaluation settings. All experiments use a 3% patch size for ImageNet and ImageNette, and 2.4% for CIFAR-10, with a fixed mask size of 56. Both strategies ensure a k-fold coverage of $k = 6$. Replicated Tiling applies repeated mask grids at the same positions, while Offset Tiling shifts the masks to achieve complementary coverage.

Replicated Tiling consistently yields higher certified robust accuracy across all models and datasets. On ImageNet, it reaches 73.7% with ResNet, outperforming Offset Tiling's 68.5%; similar improvements are seen for ViT (78.6% vs. 74.0%) and MLP (72.5% vs. 67.9%). This robustness gain arises because the fixed mask layout avoids excessive occlusion of critical regions, while Offset Tiling's shifted masks are more likely to disrupt semantically important areas.

Offset Tiling, however, achieves slightly better clean accuracy in some cases, such as ResNet on ImageNette (99.7% vs. 99.5%) and MLP on CIFAR-10 (98.0% vs. 95.7%), due to its more uniform occlusion pattern, which better preserves clean input information.

Impact of k on clean and certified robust accuracy. Table **??** analyzes the effect of varying the mask coverage parameter k on the clean and certified robust accuracy of CertMask using ViT on ImageNet. We fix the patch size to 3% and mask size to 56, and sweep k from 1 to 6 while maintaining the Offset tiling pattern.

As k increases, the certified robust accuracy generally de-

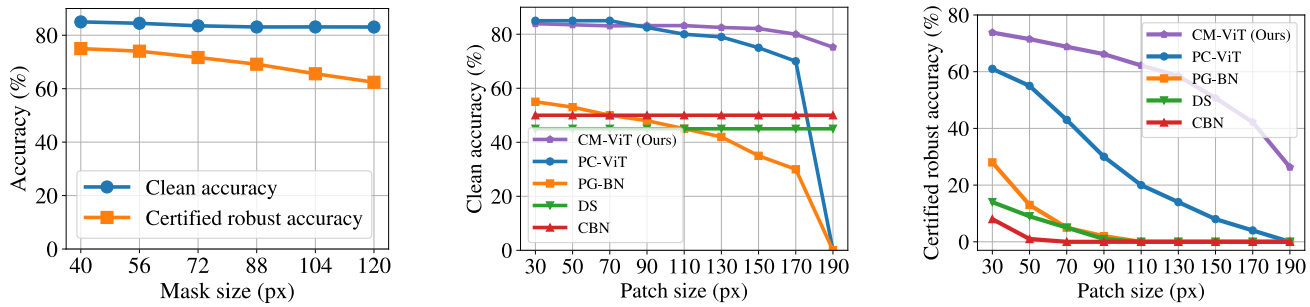

Figure 6: The effect of mask size on defense performance (left), clean accuracy under different patch sizes (middle), and certified robust accuracy under different patch sizes (right).

creases, dropping from 78.6% at $k = 1$ to 74.0% at $k = 6$. This trend reflects the tradeoff introduced by heavier mask coverage: while larger k values strengthen theoretical certification guarantees by covering each patch location more times, they also lead to increased occlusion, which can harm clean classification. Correspondingly, clean accuracy improves with larger k, increasing from 82.0% at $k = 1$ to 84.5% at $k = 6$, since higher k reduces the number of masks per image, resulting in less aggressive masking per forward pass. This observation highlights that moderate values of k (e.g., $k = 3$ or $k = 4$) may offer the best balance between certified robustness and clean performance.

Effect of Mask Size on Performance. Figure 6 shows how varying the mask size affects clean and certified robust accuracy under a fixed mask coverage $k = 6$ for ViT on ImageNet. As the mask size increases, certified robust accuracy drops steadily (e.g., from 76% at 40px to 62% at 120px), while clean accuracy remains relatively stable. This is because larger masks occlude more input area per mask, potentially hiding critical visual features needed for correct prediction. According to the theoretical bounds in Eq. 4 and Eq. 5, a larger mask size reduces the number of masks required to achieve k-fold coverage, thus lowering inference time. However, this comes at the cost of degraded robustness, suggesting a trade-off between computational efficiency and defense strength.

Performance under Varying Patch Sizes. Figures 6 illustrate the impact of increasing adversarial patch size on clean accuracy and certified robust accuracy, respectively. All evaluations are conducted on ViT with ImageNet under a fixed mask coverage of $k = 6$.

As expected, the clean and certified accuracy of all methods degrade as the patch size increases. However, CertMask (CM-ViT) demonstrates significantly slower performance degradation compared to prior defenses. For example, when the patch size increases from 30 to 130 pixels, CM-ViT maintains clean accuracy above 80% and certified accuracy around 66.8%, whereas PatchCleanser (PC-ViT) rapidly declines to nearly 0%. This robustness is attributed to CertMask's single-round, image-agnostic tiling strategy: each image is processed with a fixed and carefully optimized mask set. In contrast, PatchCleanser generates one or two adaptive masks per image, which are more sensitive to the patch size—larger patches are more likely to escape randomized coverage, leading to faster performance degradation.

Although the accuracy of CertMask does decrease as the patch size grows, the drop is substantially slower, suggesting a higher tolerance to patch perturbation. This property is particularly valuable in realistic settings where the exact patch size is unknown or may vary, demonstrating that CertMask offers not only stronger robustness but also greater practical reliability under diverse threat conditions.

Related Work

Defenses against adversarial patch attacks can be broadly classified into empirical and certified approaches. Empirical methods (Hayes 2018; Rao, Stutz, and Schiele 2020) evaluate robustness against specific attack strategies but lack formal guarantees and often fail under stronger or adaptive adversaries. In contrast, certified defenses (Xiang, Mahloujifar, and Mittal 2022; Saha et al. 2023; Li, Zhang, and Xie 2022) aim to provide provable guarantees that model predictions remain invariant under any perturbation within a defined threat model. These methods are attack-agnostic and maintain their validity even under full adversarial knowledge, offering stronger and more reliable robustness at the expense of increased computational and design complexity.

Conclusion

In this work, we present CertMask, a mathematically grounded defense framework that certifies robustness against adversarial patch attacks via optimal mask coverage. By reducing the patch-covering problem to a geometric formulation, our approach constructs a provably sufficient set of masks that guarantees full patch coverage with controllable redundancy, achieving provable robustness with significantly improved computational efficiency. Compared to prior certified defenses, CertMask achieves superior clean and certified robust accuracy with linear certification cost, enabling scalable deployment to high-resolution vision tasks. As a promising future direction, our framework may be extended to spatiotemporal settings, enabling certified defenses for video models, as well as to scenarios where the adversarial patch size is unknown.

Acknowledgements

The work is partially supported by NSF grants CPS 2521121, the Humboldt Fellowship, and the German Federal Ministry of Education and Research (BMBF) in the Course of the 6GEM Research Hub under Grant 16KISK038.

References

Brown, T. B.; Mané, D.; Roy, A.; Abadi, M.; and Gilmer, J. 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665*.

Chiang, P.-y.; Ni, R.; Abdelkader, A.; Zhu, C.; Studer, C.; and Goldstein, T. 2020. Certified defenses for adversarial patches. *arXiv preprint arXiv:2003.06693*.

Croce, F.; Andriushchenko, M.; Singh, N. D.; Flammarion, N.; and Hein, M. 2022. Sparse-rs: a versatile framework for query-efficient sparse black-box adversarial attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 6437–6445.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

fast.ai. 2020. ImageNette: A smaller subset of 10 easily classified classes from ImageNet. https://github.com/fastai/imagenette. Accessed: 2025-07-20.

Hayes, J. 2018. On visible adversarial perturbations & digital watermarking. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 1597–1604.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Karmon, D.; et al. 2018. Lavan: Localized and visible adversarial noise. In *International conference on machine learning*, 2507–2515. PMLR.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Levine, A.; and Feizi, S. 2020. (De) randomized smoothing for certifiable defense against patch attacks. *Advances in neural information processing systems*, 33: 6465–6475.

Li, J.; Zhang, H.; and Xie, C. 2022. Vip: Unified certified detection and recovery for patch attack with vision transformers. In *European Conference on Computer Vision*, 573–587. Springer.

Lin, W.-Y.; Sheikholeslami, F.; Rice, L.; Kolter, J. Z.; et al. 2021. Certified robustness against physically-realizable patch attack via randomized cropping.

McCoyd, M.; Park, W.; Chen, S.; Shah, N.; Roggenkemper, R.; Hwang, M.; Liu, J. X.; and Wagner, D. 2020. Minority reports defense: Defending against adversarial patches. In *International Conference on Applied Cryptography and Network Security*, 564–582. Springer.

Metzen, J. H.; and Yatsura, M. 2021. Efficient certified defenses against patch attacks on image classifiers. *arXiv preprint arXiv:2102.04154*.

Rao, S.; Stutz, D.; and Schiele, B. 2020. Adversarial training against location-optimized adversarial patches. In *European conference on computer vision*, 429–448. Springer.

Saha, A.; Yu, S.; Norouzzadeh, A.; Lin, W.-Y.; and Mummadi, C. K. 2023. Revisiting image classifier training for improved certified robust defense against adversarial patches. *arXiv preprint arXiv:2306.12610*.

Shi, Y.; Han, Y.; Tan, Y.-a.; and Kuang, X. 2022. Decision-based black-box attack against vision transformers via patch-wise adversarial removal. *Advances in Neural Information Processing Systems*, 35: 12921–12933.

Touvron, H.; Bojanowski, P.; Caron, M.; Cord, M.; El-Nouby, A.; Grave, E.; Izacard, G.; Joulin, A.; Synnaeve, G.; Verbeek, J.; et al. 2022. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 5314–5321.

Wang, J.; Yin, Z.; Hu, P.; Liu, A.; Tao, R.; Qin, H.; Liu, X.; and Tao, D. 2022. Defensive patches for robust recognition in the physical world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2456–2465.

Xiang, C.; Bhagoji, A. N.; Sehwag, V.; and Mittal, P. 2021. {PatchGuard}: A provably robust defense against adversarial patches via small receptive fields and masking. In *30th USENIX Security Symposium (USENIX Security 21)*, 2237–2254.

Xiang, C.; Mahloujifar, S.; and Mittal, P. 2022. {PatchCleanser}: Certifiably robust defense against adversarial patches for any image classifier. In *31st USENIX security symposium (USENIX Security 22)*, 2065–2082.

Xiao, C.; Li, B.; Zhu, J.-Y.; He, W.; Liu, M.; and Song, D. 2018. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*.

Xiao, Z.; Gao, X.; Fu, C.; Dong, Y.; Gao, W.; Zhang, X.; Zhou, J.; and Zhu, J. 2021. Improving transferability of adversarial patches on face recognition with generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11845–11854.

Yang, C.; Kortylewski, A.; Xie, C.; Cao, Y.; and Yuille, A. 2020. Patchattack: A black-box texture-based attack with reinforcement learning. In *European Conference on Computer Vision*, 681–698. Springer.

Yuan, S.; Li, H.; Han, X.; Xu, G.; Jiang, W.; Ni, T.; Zhao, Q.; and Fang, Y. 2024. Itpatch: An invisible and triggered physical adversarial patch against traffic sign recognition. *arXiv preprint arXiv:2409.12394*.

Yuan, S.; Zhang, Q.; Gao, L.; Cheng, Y.; and Song, J. 2022. Natural color fool: Towards boosting black-box unrestricted attacks. *Advances in Neural Information Processing Systems*, 35: 7546–7560.

Zhang, Z.; Yuan, B.; McCoyd, M.; and Wagner, D. 2020. Clipped bagnet: Defending against sticker attacks with clipped bag-of-features. In *2020 IEEE Security and Privacy Workshops (SPW)*, 55–61. IEEE.