CERTIFIED DEFENCES AGAINST ADVERSARIAL PATCH ATTACKS ON SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial patch attacks are an emerging security threat for real world deep learning applications. We present DEMASKED SMOOTHING, the first approach (up to our knowledge) to certify the robustness of semantic segmentation models against this threat model. Previous work on certifiably defending against patch attacks has mostly focused on image classification task and often required changes in the model architecture and additional training which is undesirable and computationally expensive. In DEMASKED SMOOTHING, any segmentation model can be applied without particular training, fine-tuning, or restriction of the architecture. Using different masking strategies, DEMASKED SMOOTHING can be applied both for certified detection and certified recovery. In extensive experiments we show that DEMASKED SMOOTHING can on average certify 64% of the pixel predictions for a 1% patch in the detection task and 48% against a 0.5% patch for the recovery task on the ADE20K dataset.

1 INTRODUCTION

Physically realizable adversarial attacks are a threat for safety-critical (semi-)autonomous systems such as self-driving cars or robots. Adversarial patches (Brown et al., 2017; Karmon et al., 2018) are the most prominent example of such an attack. Their realizability has been demonstrated repeatedly, for instance by Lee & Kolter (2019): an attacker places a printed version of an adversarial patch in the physical world to fool a deep learning system. While empirical defenses (Hayes, 2018; Naseer et al., 2019; Selvaraju et al., 2019; Wu et al., 2020) may offer robustness against known attacks, it does not provide any guarantees against unknown future attacks (Chiang et al., 2020). Therefore, certified defenses for the patch threat model, which allow guaranteed robustness against all possible attacks for the given threat model, are crucial for safety-critical applications.

Research on certifiable defenses against adversarial patches can be broadly categorized into certified recovery and certified detection. *Certified recovery* (Chiang et al., 2020; Levine & Feizi, 2020; Zhang et al., 2020; Xiang et al., 2021; Metzen & Yatsura, 2021; Lin et al., 2021; Xiang et al., 2022a; Salman et al., 2021; Chen et al., 2022) has the objective to make a correct prediction on an input even in the presence of an adversarial patch. In contrast, *certified detection* (McCoyd et al., 2020; Xiang & Mittal, 2021b; Han et al., 2021; Huang & Li, 2021) provides a weaker guarantee by only aiming at *detecting* inputs containing adversarial patches. While certified recovery is more desirable in principle, it typically comes at a high cost of reduced performance on clean data. In practice, certified detection might be preferable because it allows maintaining high clean performance. Most existing certifiable defenses against patches are focused on image classification, with the exception of DetectorGuard (Xiang & Mittal, 2021a) and ObjectSeeker (Xiang et al., 2022b) that certifiably defend against patch hiding attacks on object detectors. Moreover, existing defences are not easily applicable to arbitrary downstream models, because they assume either that the downstream model is trained explicitly for being certifiably robust (Levine & Feizi, 2020; Metzen & Yatsura, 2021), or that the model has a certain network architecture such as BagNet (Zhang et al., 2020; Metzen & Yatsura, 2021; Xiang et al., 2021) or a vision transformer (Salman et al., 2021; Huang & Li, 2021). A notable exception is PatchCleanser (Xiang et al., 2022a), which can be combined with arbitrary downstream models but is restricted to image classification.

Adversarial patch attacks were also proposed for the image segmentation problem (Nesti et al., 2022), mostly for attacking CNN-based models that use a localized receptive field (Zhao et al., 2017).

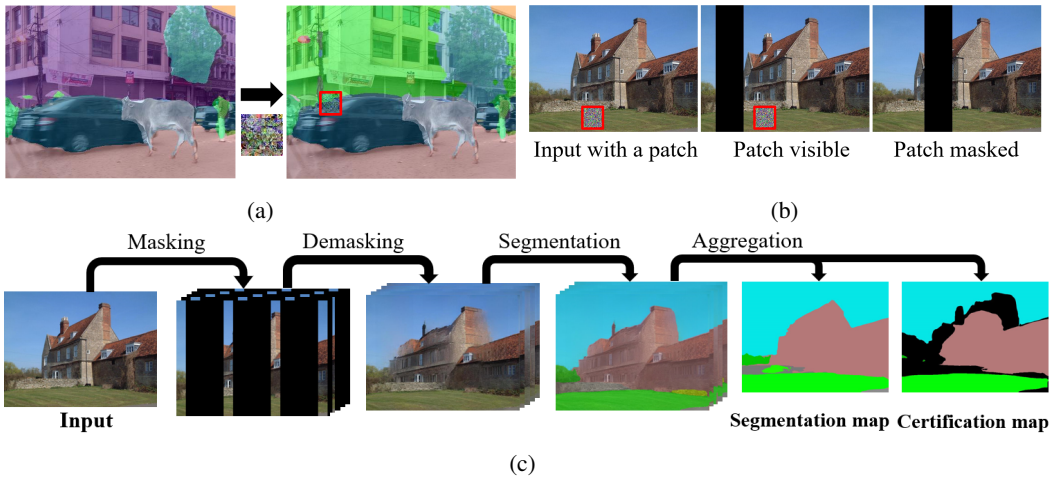


Figure 1: (a) A simple patch attack on the state-of-the-art ViT-based Senformer Bousselham et al. (2021) manages to switch the prediction for a large object. (b) Masking the patch. (c) A sketch of DEMASKED SMOOTHING for certified image segmentation. First, we generate a set of masked versions of the image such that each possible patch can only affect a certain number of masked images. Then we use image inpainting to partially recover the information lost during masking and then apply an arbitrary segmentation method. The output is obtained by aggregating the segmentations pixelwise. The masking strategy and aggregation method depend on the certification mode (detection or recovery).

However, recently self-attention based vision transformers (Dosovitskiy et al., 2021) have achieved new state-of-the-art in the image segmentation task (Liu et al., 2021; Bousselham et al., 2021). Their output may become more vulnerable to adversarial patches if they manage to manipulate the global self-attention (Lovisotto et al., 2022). We demonstrate how significant parts of the segmentation output may be affected by a small patch for the Senformer Bousselham et al. (2021) model in Figure 1a. We point out that preventive certified defences are important because newly developed attacks can immediately be used to compromise safety-critical applications unless they are properly defended.

In this work, we propose the novel framework DEMASKED SMOOTHING (Figure 1c) to obtain the first (to the best of our knowledge) certified defences against patch attacks on semantic segmentation models. Similarly to previous work (Levine & Feizi, 2020), we mask different parts of the input (Figure 1b) and provide guarantees with respect to every possible patch that is not larger than a certain pre-defined size. While prior work required the classification model to deal with such masked inputs, we leverage recent progress in image inpainting (Dong et al., 2022) to reconstruct the input *before* passing it to the downstream model. This decoupling of image demasking from the segmentation task allows us to support arbitrary downstream models. Moreover, we can leverage state of the art methods for image inpainting. We also propose different masking schemes tailored for the segmentation task that provide the dense input allowing the demasking model to understand the scene but still satisfy the guarantees with respect to the adversarial patch.

We summarize our contributions as follows:

- We propose DEMASKED SMOOTHING which is the first (to the best of our knowledge) certified recovery or certified detection based defence against adversarial patch attacks on semantic segmentation models (Section 4).
- DEMASKED SMOOTHING can do certified detection and recovery with any possible segmentation model without requiring finetuning or any other adaptation.
- We implement DEMASKED SMOOTHING, evaluate it for different certification objectives and masking schemes (Section 5). We can certify 64% of all pixels in certified detection for a 1% patch and 48% in certified recovery for a 0.5% patch for the Senformer (Bousselham et al., 2021) segmentation model on the ADE20K Zhou et al. (2017) dataset.

2 RELATED WORK

Certified recovery. The first certified recovery defence for classification models against patches was proposed by Chiang et al. (2020), who adapted interval-bound propagation (Gowal et al., 2019) to the patch threat model. Levine & Feizi (2020) proposed De-Randomized Smoothing (DRS), which provides significant accuracy improvement when compared to Chiang et al. (2020) and scales to the ImageNet dataset. In DRS, a base classifier is trained on images where everything but a small local region is masked (ablated). At inference time, a majority vote of all specified ablations is taken as the final classification. If this vote has a large enough margin to the runner-up class, the prediction cannot be shifted by any patch that does not exceed a pre-defined size. A similar approach was adopted in Randomized Cropping (Lin et al., 2021). A general drawback of these approaches is that the classifier needs to be trained to process masked/cropped inputs, which (in contrast to our work) prohibits the usage of arbitrary pretrained models. A further line of work studies network architectures that are particularly suited for certified recovery. For instance, models with small receptive fields such as *BagNets* (Brendel & Bethge, 2019) have been explored, either by combining them with some fixed postprocessing (Zhang et al., 2020; Xiang et al., 2021) or by training them end-to-end for certified recovery (Metzen & Yatsura, 2021). Salman et al. (2021) propose to apply DRS to *Vision Transfomers (ViTs)*. In contrast to the aforementioned works, our Demasked Smoothing can be applied to models with arbitrary architecture. This is a property shared with PatchCleanser (Xiang et al., 2022a), which however is limited to image classification and it is not clear how it can be extended to semantic segmentation where a class needs to be assigned to every pixel including the masked ones. Certified recovery against patches has also been extended to object detection, specifically to defend against patch hiding attacks. Two notable works in this direction are DetectorGuard (Xiang & Mittal, 2021b), an extension of PatchGuard (Xiang et al., 2021) to object detection, and ObjectSeeker (Xiang et al., 2022b). Randomized smoothing (Cohen et al., 2019) has been applied to certify semantic segmentation models against ℓ_2-norm bounded adversarial attacks (Fischer et al., 2021). However, to the best of our knowledge, no certified defence against patch attacks for semantic segmentation has been proposed so far.

Certified detection. An alternative to certified recovery is *certified detection*. Here, an adversarial patch is allowed to change the model prediction. However, if it succeeds in doing so, there is a mechanism that detects this attack certifiably with zero false negatives. Minority Reports (McCoyd et al., 2020) was the first certified detection method against patches, which is based on sliding a mask over the input in a way that ensures that there will be one mask position that completely hides the patch. PatchGuard++ (Xiang & Mittal, 2021b) is an extension of Minority Reports where the sliding mask is not applied on the input but on the feature maps of a BagNet-type feature extractor. This reduces inference time drastically since the feature extractor needs to be executed only once per input. ScaleCert (Han et al., 2021) tries to identify "superficial important neurons", which allows pruning the network in a way that the prediction needs to be made for fewer masked inputs. Lastly, PatchVeto (Huang & Li, 2021) is a recently proposed method for certified detection that is tailored towards ViT models. It implements masking by removing certain input patches of the ViT. In this work, we propose a novel method for certified detection in the semantic segmentation task that can be used for any pretrained model.

Image reconstruction. The problem of learning to reconstruct the full image from inputs where parts have been masked out was pioneered by Vincent et al. (2010). It recently attracted attention as proxy task for self-supervised pre-training, especially for the ViTs (Bao et al., 2022; He et al., 2021). Recent approaches to this problem are using Fourier convolutions (Suvorov et al., 2022) and visual transformers (Dong et al., 2022). SPG-Net (Song et al., 2018) trains a subnetwork to reconstruct the full semantic segmentation directly from the masked input as a part of the image inpainting pipeline. In this work, we use the state-of-the-art ZITS (Dong et al., 2022) inpainting method.

3 PROBLEM SETUP

3.1 SEMANTIC SEGMENTATION

In this work, we focus on the semantic segmentation task. Let $\mathcal{X}$ be a set of rectangular images. Let $x \in \mathcal{X}$ be an image with height H, width W and the number of channels C. We denote $\mathcal{Y}$ to be a finite label set. The goal is to find the segmentation map $s \in \mathcal{Y}^{H \times W}$ for x. For each pixel

$x_{i,j}$, the corresponding label $s_{i,j}$ denotes the class of the object to which $x_{i,j}$ belongs. We denote $\mathbb{S}$ to be a set of segmentation maps and $f : \mathcal{X} \to \mathbb{S}$ to be a segmentation model. Assume that we know the ground truth segmentation map $s \in \mathbb{S}$. We evaluate a model f by using *accuracy map* $M^{acc}(f, x, s) \in \{0, 1\}^{H \times W}$ such that $M^{acc}(f, x, s)_{i,j} := [f(x)_{i,j} = s_{i,j}]$, where $[P] = 1$ if $P =$ True and 0 otherwise. Let $Q(f, x, s)$ be a segmentation quality metric e.g. *global accuracy*, $Acc(f, x, s) := \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} M^{acc}(f, x, s)_{i,j}$. We discuss other metrics for the segmentation evaluation in Section 5.2.

3.2 Threat model

Let us consider an untargeted adversarial patch attack on a segmentation model. Consider an image $x \in [0, 1]^{H \times W \times C}$ and its ground truth segmentation map s. Assume that the attacker can modify an arbitrary rectangular region of the image x which has a size of $H' \times W'$. We refer to this modification as a *patch*. Let $l \in \{0,\ 1\}^{H \times W}$ be a binary mask that defines the patch location in the image in which ones denote the pixels belonging to the patch. Let $\mathcal{L}$ be a set of all possible patch locations for a given image x. Let $p \in [0, 1]^{H \times W \times C}$ be the modification itself. Then we define an operator A as $A(x,\ p,\ l) = (1 - l) \odot x + l \odot p$, where $\odot$ is element-wise product. The operator A applies the $H' \times W'$ subregion of p defined by a binary mask l to the image x while keeping the rest of the image unchanged. We denote $\mathcal{P} := [0,\ 1]^{H \times W \times C} \times \mathcal{L}$ to be a set of all possible patch configurations (p, l) that define an $H' \times W'$ patch. The goal of an attacker is to find $(p^\star,\ l^\star)$ s. t. $(p^\star,\ l^\star) = \underset{(p,\ l) \in \mathcal{P}}{\arg\min}\ Q(f, A(x, p, l), s)$

3.3 Defence objective

In this paper, we propose certified defences against patch attacks. It means that we certify against *any possible attack* from $\mathcal{P}$ including $(p^\star,\ l^\star)$. We consider two robustness objectives.

Certified recovery For a pixel $x_{i,j}$ our goal is to verify that the following statement is true

$$\forall\ (p,\ l) \in \mathcal{P} :\ f(A(x,\ p,\ l))_{i,j} = f(x)_{i,j} \tag{1}$$

Certified detection We consider a verification function v defined on $\mathcal{X}$ such that $v(x) \in \{0, 1\}^{H \times W}$. If $v(x)_{i,j} = 1$, then we claim that there is no adversarial patch in the image x or the prediction on a pixel $x_{i,j}$ is not affected by this patch. Otherwise $v(x)_{i,j} = 0$ which means that a potential patch attack was detected and we return an alert. For a pixel $x_{i,j}$ our goal is to verify that the following statement is true

$$\forall\ (p,\ l) \in \mathcal{P} : v(A(x, p, l))_{i,j} = 1 \Rightarrow f(A(x,\ p,\ l))_{i,j} = f(x)_{i,j} \tag{2}$$

The secondary objective is to keep the alert ratio as small as possible and not return an alert for the pixels that satisfy 1.

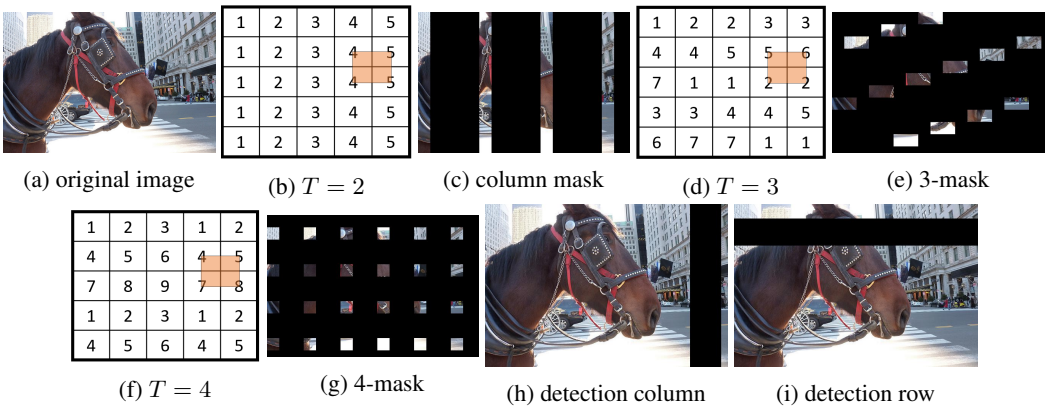
Depending on the objective our goal is to certify one of the conditions 1, 2 for each pixel $x_{i,j}$. It gives us a lower bound on attacker objective under any adversarial patch attack from $\mathcal{P}$.

4 Demasked Smoothing

Demasked Smoothing (Figure 1c) consists of several steps. First, we apply a predefined set of masks with specific properties to the input image to obtain a set of masked images. Then we reconstruct the masked regions of each image based on the available information with an inpainting model g. After that we apply a segmentation model f to the demasked results. Finally, we aggregate the segmentation outcomes and make a conclusion for the original image with respect to the statements equation 1 or equation 2.

4.1 Input masking

Motivation. Like in previous work (Section 2) we apply masking patterns to the input image and use predictions on masked images to aggregate the robust result. If an adversarial patch is completely masked, it has no effect on further processing. However, in semantic segmentation, we predict not



(a) original image (b) $T = 2$ (c) column mask (d) $T = 3$ (e) 3-mask

(f) $T = 4$ (g) 4-mask (h) detection column (i) detection row

Figure 2: examples of a mask for the column masks with $T = 2$ (b, c), 3-mask with $T = 3$ (d, e), and 4-mask with $T = 4$ (f, g) with the number of masks $K = 5, 7, 9$ respectively. The number of a block denotes in which mask it is not masked (there is only one such mask for each block). For each mask set, we show one of the locations l in which an adversarial patch affects T different maskings.

a single whole-image label like in the classification task, but a separate label for each pixel. Thus, making prediction on a masked image must allow us to predict the labels also for the masked pixels.

Preliminaries. Consider an image $x \in [0, 1]^{H \times W \times C}$. We define "$*$" to be a special masking symbol that does not correspond to any pixel value and has the property that $\forall z \in \mathbb{R} : \; z \times * = *$. Let $m \in \{*, 1\}^{H \times W}$ be a mask. We call the element-wise product $x \odot m$ a *masking* of x where a subset of pixels is hidden with * and the rest remains unchanged. We consider the threat model $\mathcal{P}$ with patches of size $H' \times W'$ (Section 3.2). We say that a mask m *covers* a patch (p, l) if $\forall\, 1 \leq i \leq H, 1 \leq j \leq W : \; (l \odot m)_{i,j} \neq 1$. It means that the patch content has no effect on the masked image i. e. $A(x,\ p,\ l) \odot m = x \odot m$. We divide the image x into a set B of non-intersecting blocks having the same size $H' \times W'$ as the adversarial patch (p, l). $B = \{b_{q,r}\}, 1 \leq q \leq \lceil H/H' \rceil$, $1 \leq r \leq \lceil W/W' \rceil$. We consider a set M of K masks, $|M| = K$. We refer to the masks in the set M by their index k as $M[k]$. To each mask $M[k]$ we assign a subset of blocks $B_k \subset B$ that are not covered in this mask while the blocks $B \setminus B_k$ are covered. We construct the masks so that each block is not covered in exactly one mask. We explain the reason for this in Section 4.2. For a mask set M we define its *strength* $T(M)$ as $T(M) = \max_{(p,l) \in \mathcal{P}} |\{m \in M \; | A(x,\ p,\ l) \odot m \neq x \odot m\}|$. $T(M)$ is the largest number of masks that can be affected by some patch. If M is defined, we refer to the value $T(M)$ as T for simplicity of notation.

Certified recovery. We define a set of column masks for which $T = 2$. We assign every k-th column to be not covered in the mask $M[k]$ (Figure 2c). Any $(p, l) \in \mathcal{P}$ can intersect at most two adjacent columns since (p, l) has the same width as a column. Thus, it can be not covered in at most two masks (Figure 2b). A similar scheme can be proposed for the rows. Due to the block size in B, the patch (p, l) cannot intersect more than four blocks at once. We define a mask set that we call *3-mask* s. t. for any four adjacent blocks two belong to the same mask (Figures 2d). Hence $T = 3$ for 3-mask. See details in Appendix A. To achieve $T = 4$ any assignment of blocks to the masks works. We consider *4-mask* that allows uniform distribution of the unmasked blocks in the image (Figure 2f).

Certified detection. We define $M_d = \{M_d[k]\}_{k=1}^{K}$ to be a set of masks s. t. for $\forall\, (p, l) \in \mathcal{P}\ \exists\, m \in M_d : A(x,\ p,\ l) \odot m = x \odot m$ i. e. for every patch exists at least one mask that covers this patch. For a patch of size $H' \times W'$ we consider $K = W - W' + 1$ masks such that the mask $M_d[k]$ masks a column of width W' starting at the horizontal position k in the image (Figure 2h). To obtain the guarantee for the same $\mathcal{P}$ with a smaller K, we consider a set of strided columns of width $W'' \geq W'$ and stride $W'' - W' + 1$ that also satisfy the condition (see the proof adapted from PatchCleanserXiang et al. (2022a) in Section B in the Appendix). A similar scheme can be proposed for the rows (Figure 2i). Alternatively, we could use a set of block masks of size $H' \times W'$. Then the number of masks grows quadratically with the image resolution. Therefore, in the experiments we focus on the column and the row masking schemes.

Let g be a demasking model, $g(x \odot m) \in [0,1]^{H \times W \times C}$. The goal of g is to make the reconstruction $g(x \odot m)$ as close as possible to the original image x. For a segmentation model f we define a *segmentation set* $S(M, x, g, f) := \{S[k] = f(g(x \odot M[k])) \mid M[k] \in M\}$.

4.2 CERTIFICATION

Certified recovery. For the threat model $\mathcal{P}$ consider a set M of K masks with some strength T. We define a function $h : \mathcal{X} \to \mathbb{S}$ that assigns a label to the pixel $x_{i,j}$ via majority voting over labels assigned at each reconstructed segmentation $s \in S$. A class for the pixel that is predicted by the largest number of segmentations is assigned. We break the ties by assigning a class with a smaller index.

Theorem 1. *If the number of masks K satisfies $K \geq 2T + 1$ and for a pixel $x_{i,j}$ we have $\forall\ S[k] \in S:\ S[k]_{i,j} = h(x)_{i,j}$ (i.e. all the votes agree), then $\forall\ (p,\ l) \in \mathcal{P}:\ h(A(x, p, l))_{i,j} = h(x)_{i,j}$*

Proof. Assume that $\exists\ (p,\ l) \in \mathcal{P}$ s. t. $h(A(x, p, l))_{i,j} \neq h(x)_{i,j}$. Let us denote $x' = A(x, p, l)$ and S' as the segmentation set for x'. Then the class $h(x)_{i,j}$ did not get the majority vote for S'. However, by definition of T we know that $(p,\ l)$ could affect at most T segmentations. Since all K segmentations of S have voted for $h(x)_{i,j}$, then at least $K - T > K/2$ of them are still voting for $h(x)_{i,j}$ in S'. Therefore $h(x')_{i,j} = h(x)_{i,j}$ □

In Section 4.1 we said that by construction each block is not covered in exactly one mask. Assume that two masks keep the same block of the image unmasked. By placing a patch so that it intersects this block, the attacker can affect both of the corresponding segmentations. Thus, they can be counted as one when computing $T(M)$. It would result in producing redundant segmentations that do not affect the certification. Therefore, we consider the case of masks with non-overlapping visible blocks in Section 4.1. Since each masking keeps approximately $1/K$ of the image pixels visible, we are interested in keeping K as small as possible to maintain more visual information for the demasking model g. Thus, for the evaluation of masking schemes with $T = 2, 3, 4$ we use the smallest possible values of $K = 5, 7, 9$ respectively (Figure 2). We observe that although for $T = 2$ we can keep a larger fraction of pixels unmasked, $T = 3, 4$ provide visible blocks that are spread in the image more uniformly. We evaluate these masking approaches in Section 5.

Certified detection. Consider $M_d = \{M_d[k]\}_{k=1}^{K}$. For a set of demasked segmentations S we define the verification map $v(x)_{i,j} := [f(x)_{i,j} = S[1]_{i,j} = \ldots = S[K]_{i,j}]$.

Theorem 2. *Assume that $v(x)_{i,j} = 1$. Then $\forall\ (p,\ l) \in \mathcal{P} : v(A(x, p, l))_{i,j} = 1 \Rightarrow f(A(x,\ p,\ l))_{i,j} = f(x)_{i,j}$*

Proof. Assume that $\exists\ (p,\ l) \in \mathcal{P}$ s. t. $v(A(x, p, l))_{i,j} = 1$ and $f(A(x,\ p,\ l))_{i,j} \neq f(x)_{i,j}$.Let us denote $x' = A(x, p, l)$ and S' as the segmentation set for x'. By definition of M_d, $\exists\ M_d[k] \in M_d$ s. t. $M_d[k]$ masks the patch $(p,\ l)$. Hence,

$$g(x \odot M_d[k]) = g(x' \odot M_d[k]),$$

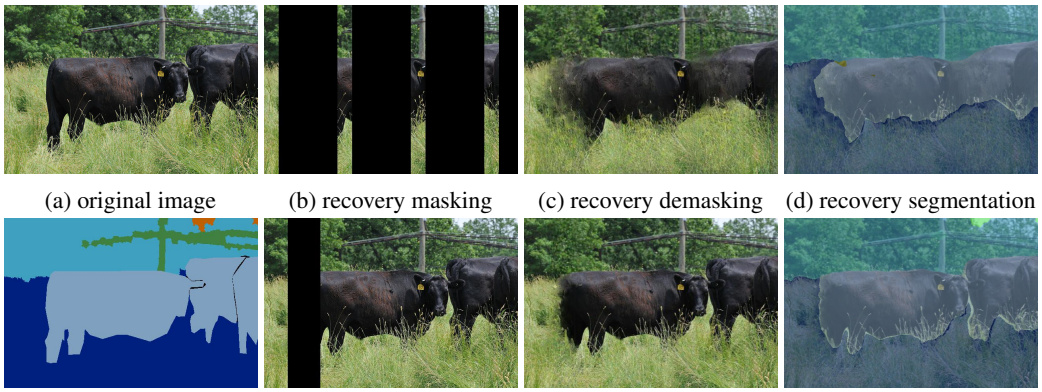
$$S[k] = f(g(x \odot M_d[k])) = f(g(x' \odot M_d[k])) = S'[k],$$

Since $v(x)_{i,j} = 1$, we have $f(x)_{i,j} = S[k]_{i,j}$. Since $v(x')_{i,j} = 1$, we have $f(x')_{i,j} = S'[k]_{i,j}$. Thus, $f(x')_{i,j} = f(x)_{i,j}$. □

For a given image x the verification map $v(x)$ is complementary to the of the model segmentation output $f(x)$ that stays unchanged. Thus, there is no drop in clean performance.

5 EXPERIMENTS

In this section, we evaluate DEMASKED SMOOTHING with the masking schemes proposed in Section 4. Certified recovery and certified detection provide certificates of different strength (Section 4) which are not comparable. We evaluate them separately for different patch sizes.



(a) original image (b) recovery masking (c) recovery demasking (d) recovery segmentation

(e) ground truth (f) detection masking (g) detection demasking (h) detection segmentation

Figure 3: Reconstructing the masked image with ZITS Dong et al. (2022) and segmenting with Senformer Bousselham et al. (2021)

5.1 Experimental Setup

We evaluate DEMASKED SMOOTHING on two challenging semantic segmentation datasets: ADE20K (Zhou et al., 2017) and COCO-Stuff-10K (Caesar et al., 2018) having 150 and 171 semantic classes respectively. Validation sets of ADE20K and COCO-Stuff-10K consist of 2000 and 1000 images respectively. For demasking we use the ZITS Dong et al. (2022) inpainting model. We use the checkpoint provided in the official paper repository [1]. The model was trained on Places2 (Zhou et al., 2016) dataset with images resized to 256×256. As a segmentation model f we use Senformer Bousselham et al. (2021), Swin Liu et al. (2021), PSPNet Zhao et al. (2017) and DeepLab v3 (Chen et al., 2018) (see Section E in the Appendix for full experimental results). We note that the first two models are based on transformers and obtain near state-of-the-art result on the considered datasets. PSPNet and DeepLab v3 are CNN-based segmentation method that we consider to demonstrate that DEMASKED SMOOTHING is not specific to transformer-based architectures and can be applied to an arbitrary segmentation model. We use the segmentation model implementations provided in the *mmsegmentation* framework Contributors (2020). For each architecture we select the best performing checkpoint. An illustration of the image reconstruction and respective Senformer segmentation can be found in Figure 3. Since evaluation on different images can be parallelized, we split the images equally among 5 Nvidia Tesla V100-32GB GPUs. The certification for the whole ADE20K validation set with ZITS and Senformer takes around 1.2 hours for certified recovery and 2 hours for certified detection (due to a larger number of masks). We point out that we did not focus on reaching the fastest possible execution time and it can be further improved by speeding up the corresponding modules. We provide more details in the Appendix.

5.2 Evaluation metrics

For both certified recovery and certified detection we produce a standard segmentation output (without any abstentions) and a corresponding certification map (Figure 4). In case of certified detection, the segmentation output remains the same as for the original segmentation model. For the certified recovery, the output is obtained by a majority vote over the segmentations of demasked images (Section 4.2) and is different from the original model output. We evaluate the mean intersection over union (mIoU) for these outputs. The certification map is obtained by assigning to each certified pixel the corresponding class from the segmentation output and assigning a special *uncertified* label to all non-certified pixels. For each image we evaluate the fraction of pixels which are certified and correct (coincide with the ground truth). %C is a mean of these fractions over all the images in the dataset. However, in semantic segmentation task, the class frequencies are usually skewed, therefore global pixel-wise accuracy alone is an insufficient metric.

[1] `https://github.com/DQiaole/ZITS_inpainting`

Table 1: The certified detection results(%) for a patch occupying no more than 1% of the image. mIoU - mean intersection over union, mR - mean recall, cmR - certified mean recall. %C - mean percentage of certified and correct pixels in the image. mIoU and mR are the clean model metrics.

dataset	segm	mask	mIoU	big		all		%C
				mR	cmR	mR	cmR	
ADE20K	Senformer	col	53.08	70.19	**59.95**	63.98	**32.90**	**64.36**
		row			51.23		26.48	59.30
	PSPNet	col	44.39	61.83	**50.02**	54.74	**26.37**	**60.57**
		row			42.44		19.88	54.62
	Swin	col	48.13	68.51	**55.45**	59.13	**29.06**	**61.44**
		row			47.21		22.04	55.93
COCO10K	Senformer	col	49.13	77.46	**63.68**	61.31	**36.05**	**53.96**
		row			56.86		31.67	50.02

(a) original image (b) recovery map (c) detection map (d) ground truth

Figure 4: Certification maps

Matching the certification map separately for each class with the ground truth segmentation of this class in the image allows us to compute the guaranteed lower bound (cTP) on the number of per-class true positive pixel predictions (TP) i.e. those that were correctly classified into this class. If a pixel was certified with a correct class, then this prediction cannot be changed by a patch (or, alternatively, the change will be detected by the verification function v in certified detection). We consider per-class *recall* $R = TP/(TP + FN)$ where FN is the number of false negative predictions for this class. $P = TP + FN$ is the total area of a given class in the ground truth and does not depend on our prediction. We can evaluate *certified recall* $cR = cTP/P$ which is a lower bound on recall R. For each class we compute its total recall in the dataset by dividing the total number of TP pixels for all images over the total P. By averaging this number among all classes we evaluate *mean recall* mR and its lower bound *certified mean recall* cmR. Evaluating lower bounds on other popular metrics such as mean precision or mIoU cannot be done this way since they depend on the upper bound on false positive (FP) predictions. For the pixels that are not certified we cannot guarantee that they will not be assigned to a certain class, therefore, a non-trivial upper bound on FP cannot be achieved. We leave this direction for future work.

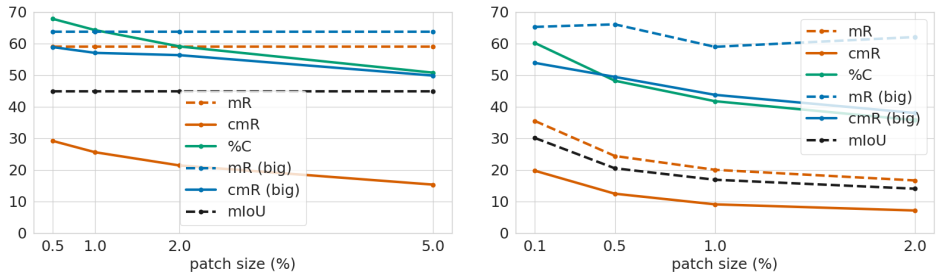
Due to our threat model, certifying small objects in the scene can be difficult because they can be partially or completely covered by an adversarial patch in a way that there is not chance to recover the prediction. To provide an additional perspective on our methods, we also evaluate their mR and cmR specifically for the "big" classes, which occupy on average more than 20% of the image in which they appear. These are, for example, road, building, train, and sky, which are important for understanding the scene. The full list of such classes for each dataset is provided in the Appendix.

5.3 Discussion

The results for certified detection and certified recovery can be found in Table 1 and Table 2 respectively. In Table 1 we see that for certified detection the column masking (Figure 2h) achieves better results than row masking in all settings. For all models we can certify around 60% of pixels on ADE20K and 50% on COCO10K. We also achieve 63.68 and 36.05 cmR for the big and all classes respectively on COCO10K with Senformer. In Table 2 we see that for certified recovery the column masking ($T = 2$, see Section 4.1) achieves the highest results in all settings. However, row

Table 2: The certified recovery results(%) against a 0.5% patch. 3-mask and 4-mask correspond to $T = 3$ and $T = 4$ respectively (Figure 2)

dataset	segm	mask	mIoU	big		all		%C
				mR	cmR	mR	cmR	
ADE20K	Senformer	col	**24.79**	**64.91**	**40.78**	**29.69**	**13.23**	**48.59**
		row	14.39	42.99	20.55	17.57	5.56	32.93
		3-mask	19.45	61.62	27.38	23.49	7.07	40.22
		4-mask	16.45	51.81	20.51	19.88	5.51	36.55
	PSPNet	col	**19.17**	**51.90**	**34.11**	**23.66**	**10.76**	**44.90**
		row	12.00	36.26	12.03	15.03	3.74	28.29
		3-mask	15.00	44.93	19.55	18.41	5.58	35.85
		4-mask	12.74	40.41	15.86	15.87	4.14	31.22
	Swin	col	**22.43**	**59.75**	**34.88**	**27.09**	**11.70**	**46.14**
		row	13.58	42.88	15.13	16.70	4.46	30.64
		3-mask	17.06	51.03	24.15	20.74	6.65	38.27
		4-mask	14.77	46.67	17.74	10.05	4.72	34.04
COCO10K	Senformer	col	**30.44**	**71.04**	**44.66**	**39.73**	**15.96**	**34.83**
		row	27.84	66.58	24.70	37.23	9.25	22.64
		3-mask	26.54	66.61	32.68	35.24	10.64	28.82
		4-mask	25.73	66.12	29.83	34.10	8.57	24.51



(a) certified detection

(b) certified recovery

Figure 5: Performance for different adversarial patch sizes evaluated on 200 ADE20K images.

masking that also satisfies $T = 2$ achieves worse results than the ones for $T = 3, 4$ on ADE20K. This suggests that not only the fraction of the image that is visible but also the mask structure affects the certification performance of DEMASKED SMOOTHING. For the recovery task, we also observe that majority voting achieves 24.79 mIoU on ADE20K and 30.44 on COCO10K compared to the original model values of 53.08 and 49.13 respectively. For Senformer on ADE20K we can on average certify almost 50% of the pixels and achieve cmR of 41.78 and 13.23 for big and all classes respectively. Figure 5 shows how the performance of DEMASKED SMOOTHING depends on the patch size. We see that certified detection metrics remain high even for a patch as big as 5% of the image surface and for the recovery they slowly deteriorate as we increase the patch size to 2%.

6 CONCLUSION

In this work, we propose the first (to the best of our knowledge) certified defences against patch attacks on segmentation models. We implement them in a framework that we call DEMASKED SMOOTHING. Due to its novel design based on masking schemes and image demasking, DEMASKED SMOOTHING is compatible with any segmentation model and can on average certify 64% of the pixel predictions for a 1% patch in the detection task and 48% against a 0.5% patch for the recovery task on the ADE20K dataset.

REFERENCES

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=p-BhZSz59o4.

Walid Bousselham, Guillaume Thibault, Lucas Pagano, Archana Machireddy, Joe Gray, Young Hwan Chang, and Xubo Song. Efficient self-ensemble for semantic segmentation, 2021. URL https://arxiv.org/abs/2111.13280.

Wieland Brendel and Matthias Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. In *International Conference on Learning Representations (ICLR)*, 2019.

Tom Brown, Dandelion Mane, Aurko Roy, Martin Abadi, and Justin Gilmer. Adversarial patch. In *Advances Neural Information Processing System (NeurIPS)*, 2017. URL https://arxiv.org/pdf/1712.09665.pdf. arXiv: 1712.09665.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

Zhaoyu Chen, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Wenqiang Zhang. Towards practical certifiable patch defense with vision transformer. *CVPR*, 2022.

Ping-yeh Chiang, Renkun Ni, Ahmed Abdelkader, Chen Zhu, Chris Studor, and Tom Goldstein. Certified defenses for adversarial patches. In *International Conference on Learning Representations (ICLR)*, 2020.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of Machine Learning Research*, pp. 1310–1320, 2019. URL http://proceedings.mlr.press/v97/cohen19c.html.

MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Marc Fischer, Maximilian Baader, and Martin T. Vechev. Scalable certified segmentation via randomized smoothing. In *ICML*, 2021.

Sven Gowal, Krishnamurthy (Dj) Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. Scalable verified training for provably robust image classification. In *International Conference on Computer Vision (ICCV)*, October 2019.

Husheng Han, Kaidi Xu, Xing Hu, Xiaobing Chen, Ling Liang, Zidong Du, Qi Guo, Yanzhi Wang, and Yunji Chen. Scalecert: Scalable certified defense against adversarial patches with sparse superficial layers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Jamie Hayes. On visible adversarial perturbations & digital watermarking. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*, 2018. URL http://openaccess.thecvf.com/content_cvpr_2018_workshops/w32/html/Hayes_On_Visible_Adversarial_CVPR_2018_paper.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. arXiv:2111.06377.

Yuheng Huang and Yuanchun Li. Zero-shot certified defense against adversarial patches with vision transformers, 2021. arXiv:2111.10481.

Danny Karmon, Daniel Zoran, and Yoav Goldberg. LaVAN: Localized and visible adversarial noise. In *International Conference on Machine Learning (ICML)*, pp. 2507–2515, 2018. URL https://proceedings.mlr.press/v80/karmon18a.html.

Mark Lee and J. Zico Kolter. On physical adversarial patches for object detection. *International Conference on Machine Learning (Workshop)*, 2019. URL http://arxiv.org/abs/1906.11897.

Alexander Levine and Soheil Feizi. (De)Randomized Smoothing for Certifiable Defense against Patch Attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Chu-Tak Li, Wan-Chi Siu, Zhi-Song Liu, Li-Wen Wang, and Daniel Pak-Kong Lun. Deepgin: Deep generative inpainting network for extreme image inpainting, 2020.

Wan-Yi Lin, Fatemeh Sheikholeslami, jinghao shi, Leslie Rice, and J Zico Kolter. Certified robustness against physically-realizable patch attack via randomized cropping, 2021. URL https://openreview.net/forum?id=vttv9ADGuWF.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision (ICCV)*, 2021.

Giulio Lovisotto, Nicole Finnie, Mauricio Munoz, Chaithanya Kumar Mummadi, and Jan Hendrik Metzen. Give me your attention: Dot-product attention considered harmful for adversarial patch robustness. *CVPR*, 2022.

Michael McCoyd, Won Park, Steven Chen, Neil Shah, Ryan Roggenkemper, Minjune Hwang, Jason Xinyu Liu, and David Wagner. Minority reports defense: Defending against adversarial patches, 2020. arXiv:2004.13799.

Jan Hendrik Metzen and Maksym Yatsura. Efficient certified defenses against patch attacks on image classifiers. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=hr-3PMvDpil.

Muzammal Naseer, Salman Khan, and Fatih Porikli. Local gradients smoothing: Defense against localized adversarial attacks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019. URL https://doi.org/10.1109/WACV.2019.00143.

Federico Nesti, Giulio Rossolini, Saasha Nair, Alessandro Biondi, and Giorgio C. Buttazzo. Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2826–2835, 2022.

Hadi Salman, Saachi Jain, Eric Wong, and Aleksander Madry. Certified patch robustness via smoothed vision transformers, 2021. arXiv:2110.07719.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, October 2019. ISSN 1573-1405.

Yuhang Song, Chao Yang, Yeji Shen, Peng Wang, Qin Huang, and C.-C. Jay Kuo. Spg-net: Segmentation prediction and guidance network for image inpainting. In *BMVC*, 2018.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor S. Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3172–3182, 2022.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(110):3371–3408, 2010. URL http://jmlr.org/papers/v11/vincent10a.html.

Tong Wu, Liang Tong, and Yevgeniy Vorobeychik. Defending against physically realizable attacks on image classification. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1909.09552.

Chong Xiang and Prateek Mittal. Detectorguard: Provably securing object detectors against localized patch hiding attacks. In *ACM Conference on Computer and Communications Security (CCS)*, 2021a.

Chong Xiang and Prateek Mittal. Patchguard++: Efficient provable attack detection against adversarial patches, 2021b. arXiv:2104.12609.

Chong Xiang, Arjun Nitin Bhagoji, Vikash Sehwag, and Prateek Mittal. Patchguard: A provably robust defense against adversarial patches via small receptive fields and masking. In *30th USENIX Security Symposium (USENIX Security)*, 2021.

Chong Xiang, Saeed Mahloujifar, and Prateek Mittal. Patchcleanser: Certifiably robust defense against adversarial patches for any image classifier. In *31st USENIX Security Symposium (USENIX Security)*, 2022a.

Chong Xiang, Alexander Valtchanov, Saeed Mahloujifar, and Prateek Mittal. Objectseeker: Certifiably robust object detection against patch hiding attacks via patch-agnostic masking, 2022b. arXiv:2202.01811.

Zhanyuan Zhang, Benson Yuan, Michael McCoyd, and David Wagner. Clipped BagNet: Defending Against Sticker Attacks with Clipped Bag-of-features. In *3rd Deep Learning and Security Workshop (DLS)*, 2020.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *Journal of Vision*, 17, 10 2016. doi: 10.1167/17.10.296.

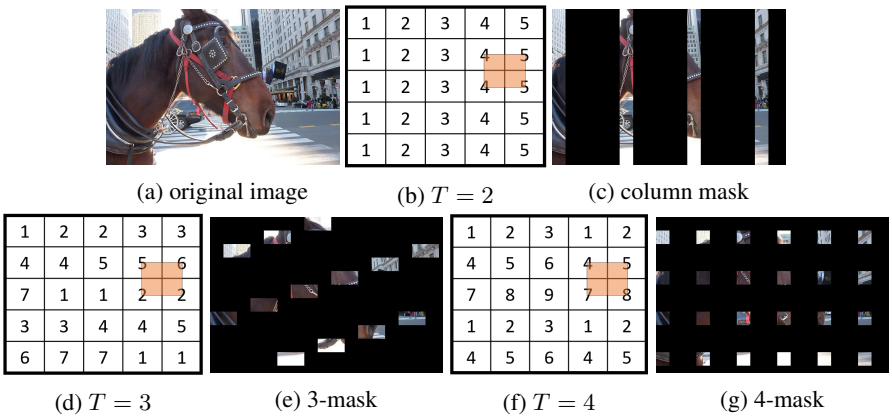
Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

A DETAILED DESCRIPTION OF MASKING STRATEGIES

In this section, we provide additional details for constructing certified recovery masks proposed in Section 4.1. We define mask sets M that satisfy different values of T. We divide the image x into a set of non-intersecting blocks $B = \{b_{q,r}\}$ of the same size as an adversarial patch, $H' \times W'$ (see Figure 6), $1 \leq q \leq \lceil H/H' \rceil$, $1 \leq r \leq \lceil W/W' \rceil$. In each mask, each of these blocks will be either masked or not masked (i. e. *visible*). Moreover, for each block there exists only one mask in which it is visible. For a set M of K masks we define the mapping $\mu_M : B \to \{1, \ldots, K\}$. If $\mu(b_{q,r}) = k$, then $b_{q,r}$ is not masked in $M[k]$. Therefore, each mask $M[k]$ is defined by a $B_k \subset B$ s. t. for $b \in B_k$ $\mu(b) = k$.

We define a set M that we call 3-mask for which $T(M) = 3$. We assign the blocks in each row to the masks as follows: $\mu(b_{1,1}) = 1$; $\mu(b_{1,2}) = \mu(b_{1,3}) = 2$; $\mu(b_{1,4}) = \mu(b_{1,5}) = 3$ and so on until we reach the end of the row. If we finish the first row with the value k, then we start the second row as follows $\mu(b_{2,1}) = \mu(b_{2,2}) = k + 1$; $\mu(b_{2,3}) = \mu(b_{2,4}) = k + 2$: .... If we finish the second row on n, we start the third row similarly to the first: $\mu(b_{3,1}) = n + 1$; $\mu(b_{3,2}) = \mu(b_{3,3}) = n + 2$; ... When we reach the number K, we start from 1 again (Figure 6d). Due to the block size, the patch cannot intersect more than four blocks at once. Our parity-alternating block sequence ensures that in any such intersection of four blocks either the top ones or the bottom ones will belong to the same masking, so at most three different maskings can be affected.

We define a set M that we call 4-mask for which $T(M) = 4$. Due to our block size any assignment of masks will work because the patch cannot intersect more than four blocks. We consider the one that allows uniform distribution of the unmasked blocks (Figure 6g). We point out that for the described methods each masking keeps approximately $1/K$ of the pixels visible and the unmasked regions are uniformly distributed in the image. This means that for any masked pixel there exists an unmasked region located close enough to this pixel. It is the core difference between our masks and the ones proposed for certified classification such as block or column smoothing Levine & Feizi (2020). It was observed that the image demasking is facilitated when the visible regions are uniformly spread in the masked image He et al. (2021).



(a) original image (b) $T = 2$ (c) column mask

(d) $T = 3$ (e) 3-mask (f) $T = 4$ (g) 4-mask

Figure 6: (a) examples of a mask for the column masks with $T = 2$ (b, c), 3-mask with $T = 3$ (d, e), and 4-mask with $T = 4$ (f, g) with the number of masks $K = 5, 7, 9$ respectively. The number of a block denotes in which mask it is not masked (there is only one such mask for each block). For each mask set, we show one of the locations l in which an adversarial patch affects T different maskings.

B CERTIFIED DETECTION MASK SIZE (SECTION 4.1)

Lemma 1. *Consider an image of the size $H \times W$. Let $H' \times W'$ be a fixed adversarial patch size. Let $M^d(K, \mathcal{L})$ be a set of masks where each mask is masking an $H \times W''$ vertical column, $W'' \geq W'$. Let the stride between the columns in two adjacent masks be $W'' - W' + 1$. Then for any location $l \in \mathcal{L}$ of the patch, there exists a mask that covers it completely.*

Proof. (Adapted from the proof of Lemma 4 in PatchCleanser Xiang et al. (2022a)). Without loss of generality, we consider the first two adjacent column masks. The first one covers the columns from 1 to W''. The second mask covers the columns from $1 + (W'' - W' + 1) = W'' - W' + 2$ to $(W'' - W' + 2) + (W'' - 1) = 2W'' - W' + 1$ (See Figure 7). Now consider an adversarial patch of size $H' \times W'$. Let us find the smallest possible start index of this patch so that it does not get covered by the first mask. For that it should be visible at the column $W'' + 1$ and, therefore, start at the column with index not smaller than $(W'' + 1) - W' + 1 = W'' - W' + 2$. However, it is the same column in which second mask starts. Therefore, given that $W'' \geq W'$ we have that the patch is completely masked by the second mask. Then for a patch which is only partially masked by the second mask from the left we use an analogous argument to show that it is completely masked by the third mask and so on. □

C TEST-TIME INPUT CERTIFICATION

In this section, we discuss how certified recovery (Theorem 1) can be applied to guaranteed verification of the robustness on a test image. We also discuss how robustness guarantees for the test-time images can be evaluated by using a dataset of clean images such as ADE20K (Zhou et al., 2017) or COCO-Stuff-10K (Caesar et al., 2018).

C.1 TEST-TIME CERTIFIED RECOVERY

Let x' be a test-time input which can be either a clean image or an image attacked with an adversarial patch. We know that there exists a clean image x corresponding to x' which removes the patch if it is present. We have either $x' = x$ or $x' \in A(x)$, where $A(x) := \{A(x, p, l) \mid (p, l) \in \mathcal{P}\}$. However, at test time we do not have access to the clean image x.

Our goal is to certify that for our segmentation model h and a pixel $x_{i,j}$ we have $h(x')_{i,j} = h(x)_{i,j}$. We can achieve this result by applying the recovery certification (Theorem 1) to the test-time image. It allows us to verify whether $\forall\ (p,\ l) \in \mathcal{P}\ :\ h(A(x', p, l))_{i,j} = h(x')_{i,j}$. We also know that if $x' \in A(x)$, then $x \in A(x')$ (Figure 8a). Indeed, if x' is only different from x by one patch, then x can be be obtained from x' by removing this patch. Therefore, by obtaining the guarantee for $A(x')$, we implicitly obtain the guarantee also for the image x even though we do not have direct access to it.

We note that this test-time guarantee is only possible for certified recovery. In certified detection, we would need to evaluate the verification function v (Theorem 2) for both the clean image x and the attacked image x' to obtain the result. This cannot be done if x is implicit.

C.2 ROBUSTNESS GUARANTEES EVALUATION

The typical certified robust error for a given test data set (and pixel (i, j) in the segmentation case) is an estimate for

$$\mathbb{E}_{X \sim D}\big[\max_{(p,l) \in P} \mathbb{1}_{h(A(X,p,l))_{i,j} \neq h(X)_{i,j}}\big],$$

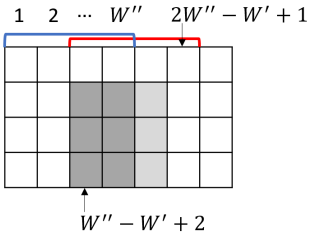


Figure 7: The masked columns of the first two adjacent masks (blue for the first one and red for the second one). If the patch is not completely masked by the first mask, it should be visible at the column $W'' + 1$ (the masked part of the patch is dark-grey and the visible part is in light-grey). However then the patch will be completely masked by the second mask.

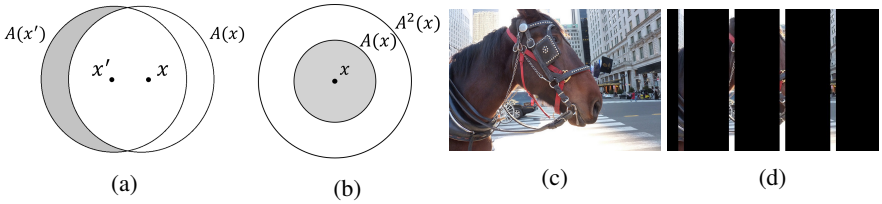


Figure 8: (a) certified inference; (b) double adversarial neighbourhood; (c) original image (d) certification against two patches

Table 3: Inference recovery robustness estimate. To illustrate our point we certify an example for a 0.1% patch.

dataset	segm	mask	mIoU	big		all		%C
				mR	cmR	mR	cmR	
ADE20K	Senformer	col	19.73	36.95	16.64	24.23	9.24	41.96
COCO10K			26.36	69.63	35.34	34.92	11.13	28.17

where D is the data generating probability measure and we assume that our test set to be an i.i.d. sample of it. This is the expected robust error (worst case over our threat model P for clean inputs) for a given pixel (i, j). Using the test sample to get an estimate of this quantity, we get a probabilistic guarantee that the corresponding pixel (i, j) of a new *clean* test sample x' drawn i.i.d. from P will have its whole "patch"-neighborhood certified.

However, more important for a practical security analysis is that we can certify a given instance, which can be even potentially adversarially perturbed. Formally, this means that for an input $z \in A(x)$, where $x \sim P$ is an unknown sample from P, that we guarantee

$$\forall (p, l) \in P : h(A(z, p, l))_{i,j} = h(z)_{i,j},$$

and as $x \in A(z, p, l)$ this implies that we certify that the pixel (i, j) of the potentially manipulated image is classified the same as pixel (i, j) of the unperturbed image x.

However, it is now tricky to get even a probabilistic estimate of the quantity

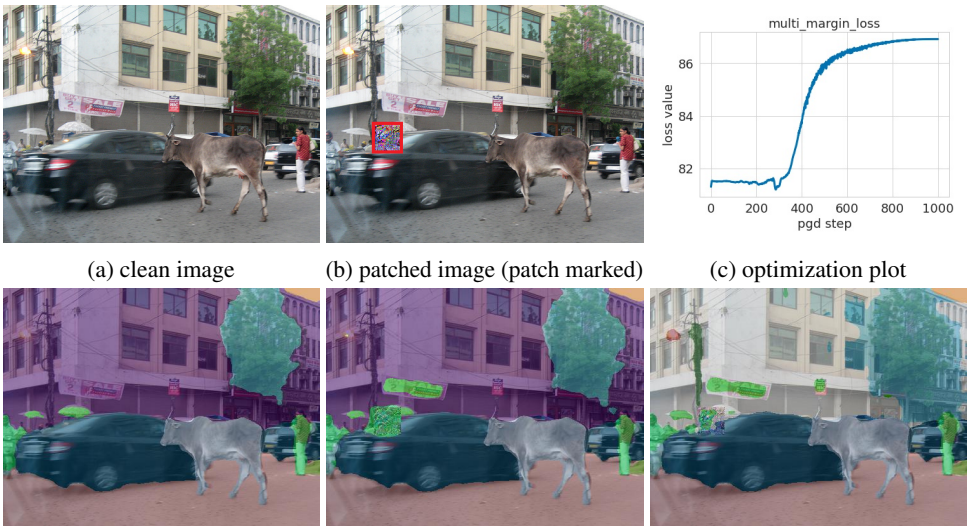
$$\mathbb{E}_{x \sim D} \max_{(p,l) \in P} [\max_{(q,m) \in P} \mathbb{1}_{h(A(A(x,p,l),q,m))_{i,j} = h(A(x,p,l))_{i,j}}],$$

as the outer maximization process cannot be simply simulated by doing adversarial patch attacks on a clean test dataset.

We propose a way to evaluate a guaranteed lower bound on the fraction of certified test-time inputs by using a dataset of clean images. Instead of considering a standard one-patch neighbourhood $A(x)$ defined by our threat model (Section 3.2), we propose to consider a neighbourhood $A^2(x)$ of two independent patches (Figure 8b). $A^2(x)$ contains all the images $x' \in A(x)$ as well as their respective patch neighbourhoods $A(x')$. Therefore, by verifying that $\forall\ (p_1,\ l_1), (p_2,\ l_2) \in \mathcal{P}\ :\ h(A(A(x,\ p_1,\ l_1),\ p_2,\ l_2))_{i,j} = h(x)_{i,j}$, we guarantee that $\forall\ x' \in A(x)\ \forall\ (p,\ l) \in \mathcal{P}\ :\ h(A(x',\ p,\ l))_{i,j} = h(x')_{i,j}$.

We note that corresponding reasoning could be applied to certification in ℓ_p models. Then $A^2(x)$ would correspond to doubling the radius of the ϵ-ball instead of adding a second patch.

Note that Theorem 1 can be directly extended to a threat model of N patches. In the worst case each of the N patches can affect T different maskings. Therefore, we need to change the condition of Theorem 1 to $K \geq 2NT + 1$. We apply the described method to evaluating the test-time certification guarantees for a toy example of a 0.1% patch in Table 3. We also illustrate how a column mask looks in this case in Figure 8.



(a) clean image (b) patched image (patch marked) (c) optimization plot

(d) clean segmentation (e) 300 PGD steps (f) 1000 PGD steps

Figure 9: Patch attack illustration. A patch occupying 1% of the image surface changes the prediction from "building-other" (purple) to "wall-other" (grey) for a large part of the image.

D REAL ADVERSARIAL PATCH

In this section, we demonstrate an example of a real adversarial patch for a semantic segmentation model similar to the one illustrated in the Figure 1a and show how it is handled by our certified defences. We illustrate it for a Senformer (Bousselham et al., 2021) model on one of the images from the COCO-Stuff-10K (Caesar et al., 2018) dataset.

D.1 PATCH OPTIMIZATION

We set the patch size to 1% of the image surface. We select a fixed position for a patch on the rear window of a car (Figure 9a). For each pixel we extract a list of predicted logits corresponding to each class and apply multi-margin loss with respect to the ground truth label of the respective pixel. We use random patch initialization without restarts. As an optimizer we use projected gradient descent (PGD) with 1000 steps and initial step size of 0.01. We use cosine step size schedule and momentum for the gradient with the rate of 0.9. The optimization plot and the patch efficiency at different iterations of the PGD are illustrated in the Figure 9.

D.2 CERTIFIED RECOVERY

We denote the original image as x and the patched image as x'. The voting-based segmentation function h (Section 4.2) provides the majority-vote prediction $h(x)$ and the corresponding certification map which shows the pixels where all the votes agree. In Figure we see that a major part of the building is certified which means that this prediction cannot be affected by an adversarial patch. Figure demonstrates $h(x')$ which correctly segments the building region in presence of an adversarial patch that fools the original model.

D.3 CERTIFIED DETECTION

We perform our analysis by evaluating the verification map v (Section 4.2) for the original image x and for the patched image x'. We see that in $v(x)$ a major part of the building is certified i. e. for a large part of pixels $x_{i,j}$ that belong to the building we have $v(x)_{i,j} = 1$. However, $v(x')_{i,j} = 0$ for all of the building pixels. It means that we have detected that the prediction on this input is potentially affected by an adversarial patch.

(a) segmentation of the original image, $h(x)$ (b) certification map of the original image x (c) segmentation of the patched image, $h(x')$ (d) certification map of the patched image x'

Figure 10: Certified recovery for a 1 % patch used in the attack. The majority vote function h recovers the prediction in presence of an adversarial patch that fools the undefended model. The segmentation for the original and patched image in (a) and (c) are the same for the regions certified in the certification maps (b) or (d). The certification maps (b) and (d) are also almost the same.

(a) $f(x)$ (original) (b) $v(x)$ (original verified) (c) $v(x')$ (patched verified) (d) $f(x')$ (patched)

Figure 11: f is a segmentation model (Senformer Bousselham et al. (2021)) and v is the verification function (Section 4.2). For an attacked image x' $v(x')$ detects the region of $f(x')$ which was (potentially) affected by an adversarial patch.

E Additional experiments on COCO10k

In this section, we provide the results for PSPNet (Zhao et al., 2017) and DeepLab v3 (Chen et al., 2018) on COCO10k (Caesar et al., 2018) dataset. We evaluate the models for certified detection in Table 4 and for certified recovery in Table 5.

Table 4: The certified detection results(%) for a patch occupying no more than 1% of the image. mIoU - mean intersection over union, mR - mean recall, cmR - certified mean recall. %C - mean percentage of certified and correct pixels in the image.

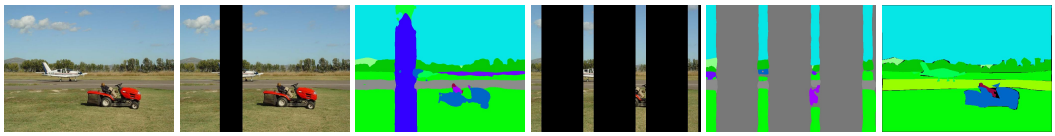
dataset	segm	mask	mIoU	big		all		%C
				mR	cmR	mR	cmR	
ADE20K	Senformer	col	53.08	70.19	**59.95**	63.98	**32.90**	**64.36**
		row			51.23		26.48	59.30
	PSPNet	col	44.39	61.83	**50.02**	54.74	**26.37**	**60.57**
		row			42.44		19.88	54.62
	Swin	col	48.13	68.51	**55.45**	59.13	**29.06**	**61.44**
		row			47.21		22.04	55.93
COCO10K	Senformer	col	49.13	77.46	**63.68**	61.31	**36.05**	**53.96**
		row			56.86		31.67	50.02
	PSPNet	col	37.76	71.71	**56.86**	49.65	**26.80**	**47.61**
		row			51.05		23.51	43.40
	DeepLab v3	col	37.81	72.52	**56.54**	49.98	**26.86**	**47.17**
		row			50.51		23.89	43.19

Table 5: The certified recovery results(%) against a 0.5% patch. 3-mask and 4-mask correspond to $T = 3$ and $T = 4$ respectively (Figure 2)

dataset	segm	mask	mIoU	big		all		%C
				mR	cmR	mR	cmR	
ADE20K	Senformer	col	**24.79**	**64.91**	**40.78**	**29.69**	**13.23**	**48.59**
		row	14.39	42.99	20.55	17.57	5.56	32.93
		3-mask	19.45	61.62	27.38	23.49	7.07	40.22
		4-mask	16.45	51.81	20.51	19.88	5.51	36.55
	PSPNet	col	**19.17**	**51.90**	**34.11**	**23.66**	**10.76**	**44.90**
		row	12.00	36.26	12.03	15.03	3.74	28.29
		3-mask	15.00	44.93	19.55	18.41	5.58	35.85
		4-mask	12.74	40.41	15.86	15.87	4.14	31.22
	Swin	col	**22.43**	**59.75**	**34.88**	**27.09**	**11.70**	**46.14**
		row	13.58	42.88	15.13	16.70	4.46	30.64
		3-mask	17.06	51.03	24.15	20.74	6.65	38.27
		4-mask	14.77	46.67	17.74	10.05	4.72	34.04
COCO10K	Senformer	col	**30.44**	**71.04**	**44.66**	**39.73**	**15.96**	**34.83**
		row	27.84	66.58	24.70	37.23	9.25	22.64
		3-mask	26.54	66.61	32.68	35.24	10.64	28.82
		4-mask	25.73	66.12	29.83	34.10	8.57	24.51
	PSPNet	col	**21.94**	**61.56**	**36.67**	**29.94**	**11.13**	**29.51**
		row	18.87	58.04	20.90	26.16	6.14	19.31
		3-mask	18.82	59.26	29.00	25.85	7.56	25.21
		4-mask	17.46	58.47	23.63	24.35	5.51	20.36
	DeepLab v3	col	**23.12**	**62.60**	**33.84**	**31.59**	**11.55**	**28.71**
		row	20.04	55.71	17.80	27.89	6.28	17.04
		3-mask	20.14	58.02	27.14	27.82	8.05	24.30
		4-mask	19.35	58.22	22.01	26.74	5.79	19.38

F INPAINTING ABLATION STUDIES

We perform ablation studies with respect to the demasking step. The results are in Table 6. Figure 12 provides additional illustrations. As can be seen from the results, our method heavily benefits from having available stronger inpainting models that allow achieving better clean and certified accuracy. We consider this property actually as a strength of our method since it will automatically benefit from future research and developments of stronger inpainting methods. For certified recovery, we also compare it to GIN Li et al. (2020) based on a generative model that we trained on ADE20K (without using style losses based on ImageNet trained VGG). The results are in Table 7. Illustrations can be found in Figure 13.



(a) original (b) detection col (c) segmentation (d) recovery col (e) segmentation (f) ground truth

Figure 12: Results without image demasking. The solid color inpainting is treated as a separate object in the scene because we need to classify every pixel in semantic segmentation task. Therefore, it is hard to achieve a situation where all the demasked segmentation agree on some pixel which is represented in the Table 6.

Table 6: Comparison for demasked smoothing with and without demasking step. mIoU - mean intersection over union, mR - mean recall, cmR - certified mean recall. %C - mean percentage of certified and correct pixels in the image. We use Swin model on 200 ADE20K images with column masking for certified detection and certified recovery. We compare masking the columns with solid black color without demasking to ZITS demasking.

mode	patch size	demasking	mIoU	big		all		%C
				mR	cmR	mR	cmR	
detection	1.0%	✓	38.56	67.25	**58.85**	53.37	**23.35**	**62.89**
		✗			19.49		3.09	21.19
recovery	0.5%	✓	**19.09**	**66.03**	**52.71**	**23.02**	**12.66**	**47.05**
		✗	1.10	15.09	7.71	1.79	0.72	18.59

Table 7: Comparison of our two demasking methods: ZITS and GIN. mIoU - mean intersection over union, mR - mean recall, cmR - certified mean recall. %C - mean percentage of certified and correct pixels in the image. We use Swin model on 200 ADE20K images.

demasking	trained on	mIoU	big		all		%C
			mR	cmR	mR	cmR	
ZITS (Dong et al., 2022)	Places2	**19.09**	**66.03**	**52.71**	**23.02**	**12.66**	**47.05**
GIN (Li et al., 2020)	ADE20K	5.46	32.27	19.05	7.62	3.52	32.08

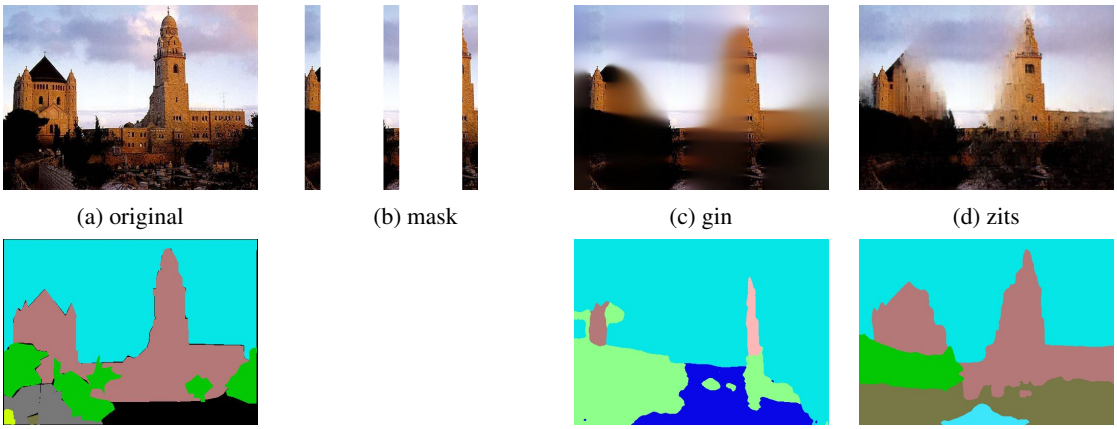
G Comparison to simplified Derandomized Smoothing

Derandomized Smoothing (DRS) Levine & Feizi (2020) was proposed for certified recovery, therefore in this section we focus on this task. Direct adaptation of derandomized smoothing to semantic segmentation task requires training a model that is able to predict the full image segmentation from a small visible region. Since it is not immediately clear to us what architectural design and training procedure would be needed to train such a model, we consider a simplified version of DRS that we call DRS-S. In this version, we consider an off-the-shelf semantic segmentation model and evaluate how it performs with column masking from DRS. Therefore, we do not encode the masked regions with the special 'NULL' value like in DRS but use black color instead. That is because an off-the-shelf model cannot work with 'NULL' values.

We run our experiments on ADE20K dataset. We consider the DRS parameters from the recent SOTA version of Derandomized Smoothing by Salman et al. Salman et al. (2021). They use column width $b = 19$ and stride $s = 10$ for certified classification of 224x224 ImageNet images. To account for the fact that ADE20K images have larger resolution than ImageNet, we scale the parameters to column width $b = 42$ and stride $s = 22$. To make the comparison consistent with the rest of our results, we use the patch occupying 0.5% of the image.

From Table 8 we can see that DRS-S performs poorly on semantic segmentation task. The reason for that is illustrated in Figure 14. Processing the column region in 14c would probably be sufficient for a classification model to classify the image into the class "house". But it is clearly not sufficient to reconstruct the whole segmentation map 14e as can be seen in the Figure 14g. Whether doing this would be possible with a model specifically trained to reconstruct the segmentation map from a very small visible region is an open research question (up to our knowledge).

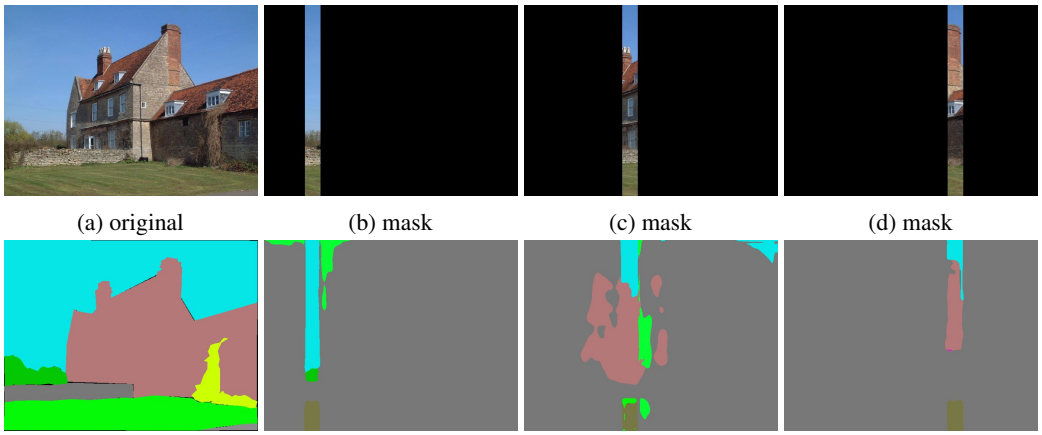
We point out that the value %C of certified and correctly classified pixels in the Table 8 is still surprisingly high for DRS-S compared to other metircs. We attribute this to the fact that the solid black regions are usually treated as a wall by the segmentation model, therefore the images are usually segmented as a wall by the DRS majority voting. And the wall is a common part of both indoor and outdoor scenes in ADE20K as can be implied from the Table 9 of "big" ADE20K classes. Therefore, always classifying the output as a wall provides a decent fraction of correctly classified pixels because of the skewed classes.



(a) original (b) mask (c) gin (d) zits

(e) ground truth (f) gin segm (g) zits segm

Figure 13: Comparison between GIN and ZITS inpainting



(a) original (b) mask (c) mask (d) mask

(e) ground truth (f) segmentation (g) segmentation (h) segmentation

Figure 14

However, to provide a better comparison with DRS, we emulate the model which is able to reconstruct the whole segmentation map from the column masking proposed in DRS. We do this by applying the demasking approach proposed in this work. We first try to reconstruct the whole image from one column and then segment it with an off-the-shelf model as we did with the masks proposed in this paper. We call this approach DRS-E and the results can be found in Table 8.

H A LIST OF BIG CLASSES

In Section 5.2 we suggest another perspective on the evaluation of our DEMASKED SMOOTHING by specifically considering its performance on "big" semantic classes. The object of these classes occupy on average more than 20% of the images in which they appear. Correctly segmenting these classes is important for understanding the scene. In Tables and we provide the full list of such classes in ADE20K (Zhou et al., 2017) and COCO-Stuff-10K (Caesar et al., 2018) together with the average fraction of pixels that they occupy in the images in which they are present. We point out that for COCO-Stuff-10K some typically smaller classes such as "sandwich" or "fruit" get included in the list of big classes because of the macro-scale images in which they occupy a big part of the scene.

Table 8: Comparison of our method with simplified Derandomized Smoothing Levine & Feizi (2020). mIoU - mean intersection over union, mR - mean recall, cmR - certified mean recall. %C - mean percentage of certified and correct pixels in the image. We use Swin model on 200 ADE20K images.

method	mIoU	big		all		%C
		mR	cmR	mR	cmR	
Demasked (our)	**19.09**	**66.03**	**52.71**	**23.02**	**12.66**	**47.05**
DRS-S	0.42	11.35	9.08	1.04	0.83	28.01
DRS-E	9.12	54.67	41.78	11.04	7.86	45.03

Table 9: The list of 19 "big" classes for ADE20K (Zhou et al., 2017) (out of 150 classes in total) with their average fraction of occupied pixels in the images where they are present (%) and index in the list of dataset classes. We define a class to be "big" if it occupies on average more than 20% of the pixels in the images in which this class appears.

#	index	name	fraction	#	index	name	fraction
1	0	wall	25.88	11	79	hovel	25.93
2	1	building	32.36	12	88	booth	23.91
3	2	sky	21.54	13	96	escalator	20.96
4	7	bed	21.25	14	103	ship	26.81
5	21	water	22.10	15	104	fountain	28.81
6	29	field	22.97	16	107	washer	22.07
7	46	sand	21.22	17	109	swimming pool	28.87
8	48	skyscraper	42.92	18	114	tent	34.57
9	54	runway	28.05	19	128	lake	34.57
10	55	case	37.57				

I DEMASKED SMOOTHING VISUALIZATION

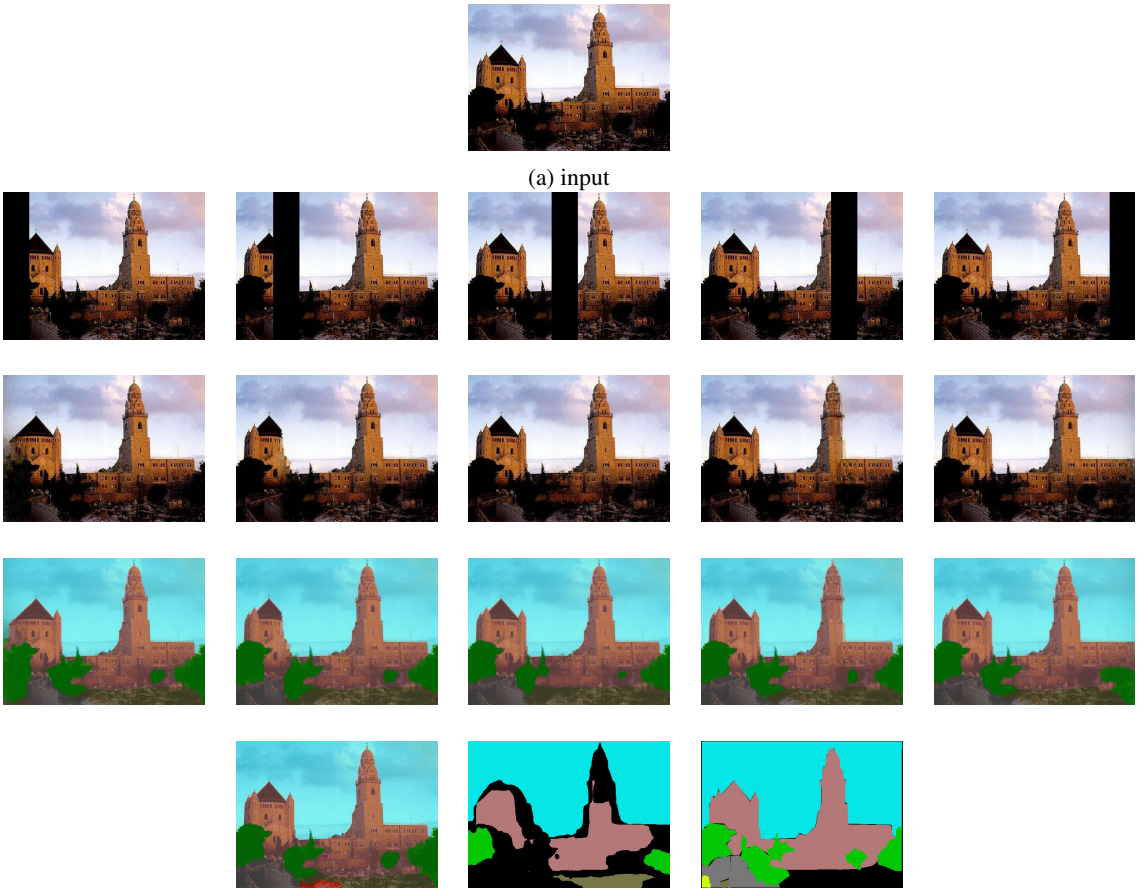
In this section, we provide additional illustrations of our method (Figures 15, 16, 17, 18). Similarly to the Tables 1 and 2 we certify against a 1% patch for the detection task and against 0.5% patch for the recovery task. For each mask type we illustrate all the stages summarized in the Figure 1c. We also provide examples of certification maps for certified recovery and certified detection with different images (Figure 19, 20).

J USED DATA

In this work, we only use the datasets published under formal licenses: ADE20K (Zhou et al., 2017) and COCO-Stuff-10K (Caesar et al., 2018). To the best of our knowledge, data used in this project do not contain any personally identifiable information or offensive content. The models ZITS (Dong et al., 2022) and Swin (Liu et al., 2021) are published under Apache-2.0 license. The text of the license for PSPNet (Zhao et al., 2017) can be found here: `https://github.com/hszhao/PSPNet/blob/master/LICENSE`

Table 10: The list of 21 "big" classes for COCO-Stuff-10K (Caesar et al., 2018) (out of 171 classes in total) with their average fraction of occupied pixels in the images where they are present (%) and index in the list of dataset classes. We define a class to be "big" if it occupies on average more than 20% of the pixels in the images in which this class appears.

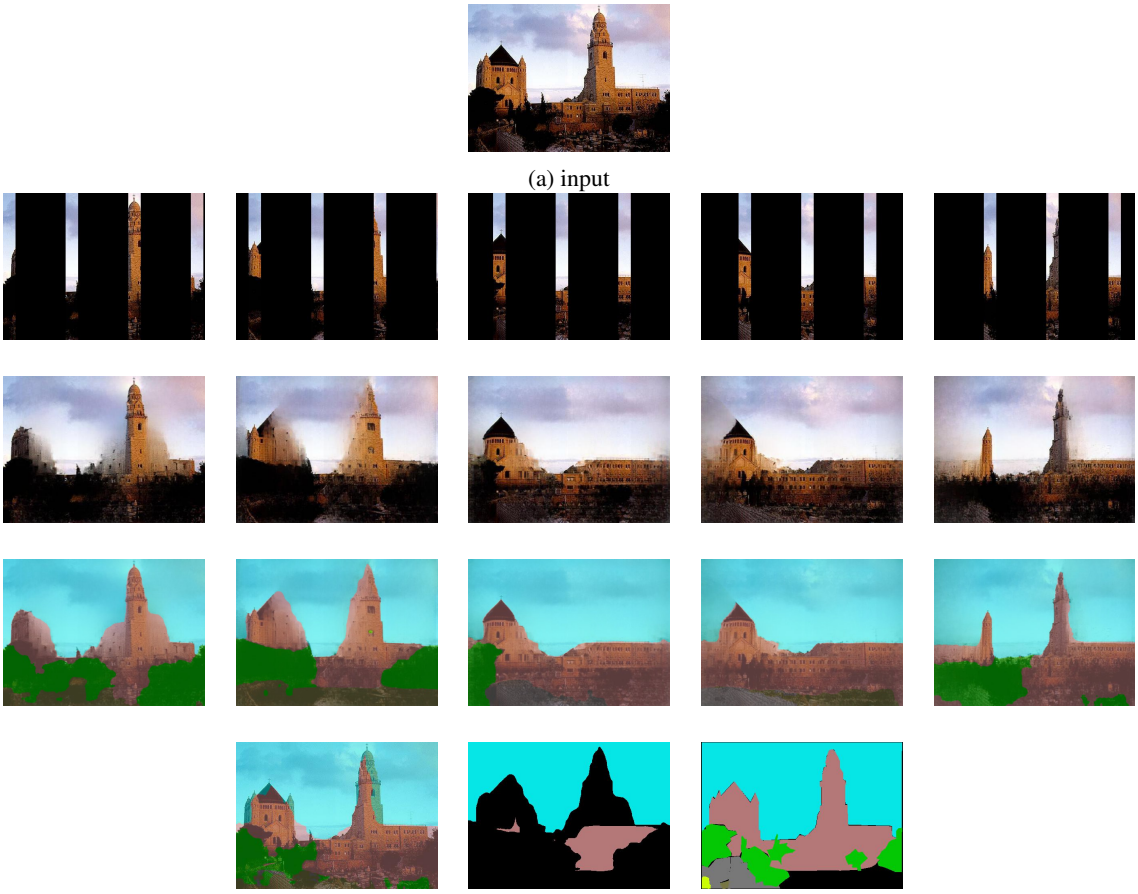
#	index	name	fraction	#	index	name	fraction
1	6	bus	21.46	11	105	floor-stone	20.10
2	7	train	23.11	12	111	fruit	20.48
3	20	cow	24.17	13	113	grass	23.25
4	21	elephant	28.50	14	134	playingfield	38.64
5	49	sandwich	23.99	15	137	river	40.01
6	51	broccoli	20.18	16	143	sand	26.37
7	54	pizza	25.86	17	144	sea	36.51
8	60	bed	36.86	18	146	sky-other	22.94
9	61	dining table	21.71	19	148	snow	51.60
10	95	clouds	24.11	20	159	vegetable	20.35
				21	167	water-other	21.67



(a) input

(b) segmentation (c) certification (d) ground truth

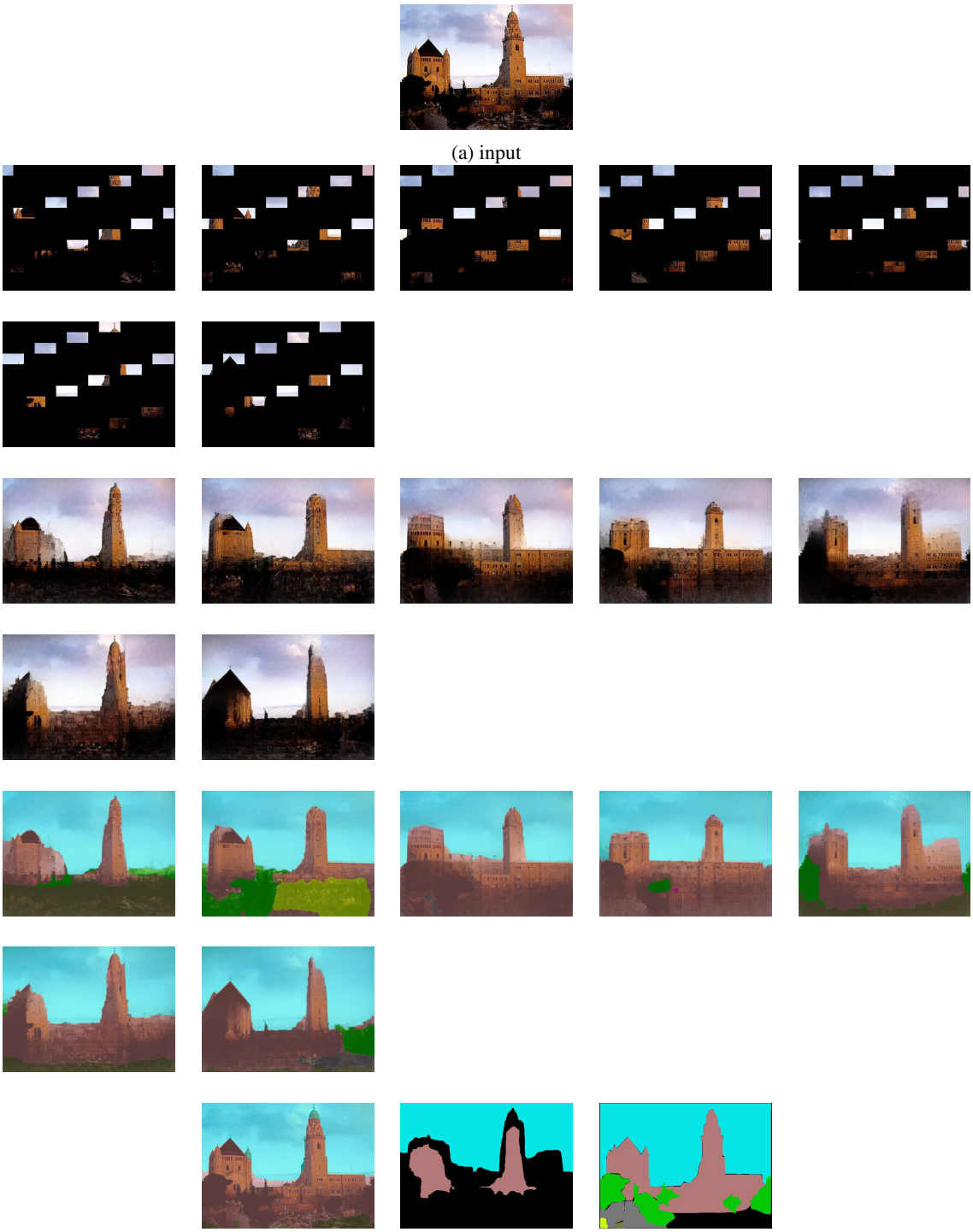
Figure 15: DEMASKED SMOOTHING detection column masking illustration for an image from ADE20K (Zhou et al., 2017) with ZITS (Dong et al., 2022) and Senformer (Bousselham et al., 2021). We illustrate five masks out of twenty.



(a) input

(b) segmentation (c) certification (d) ground truth

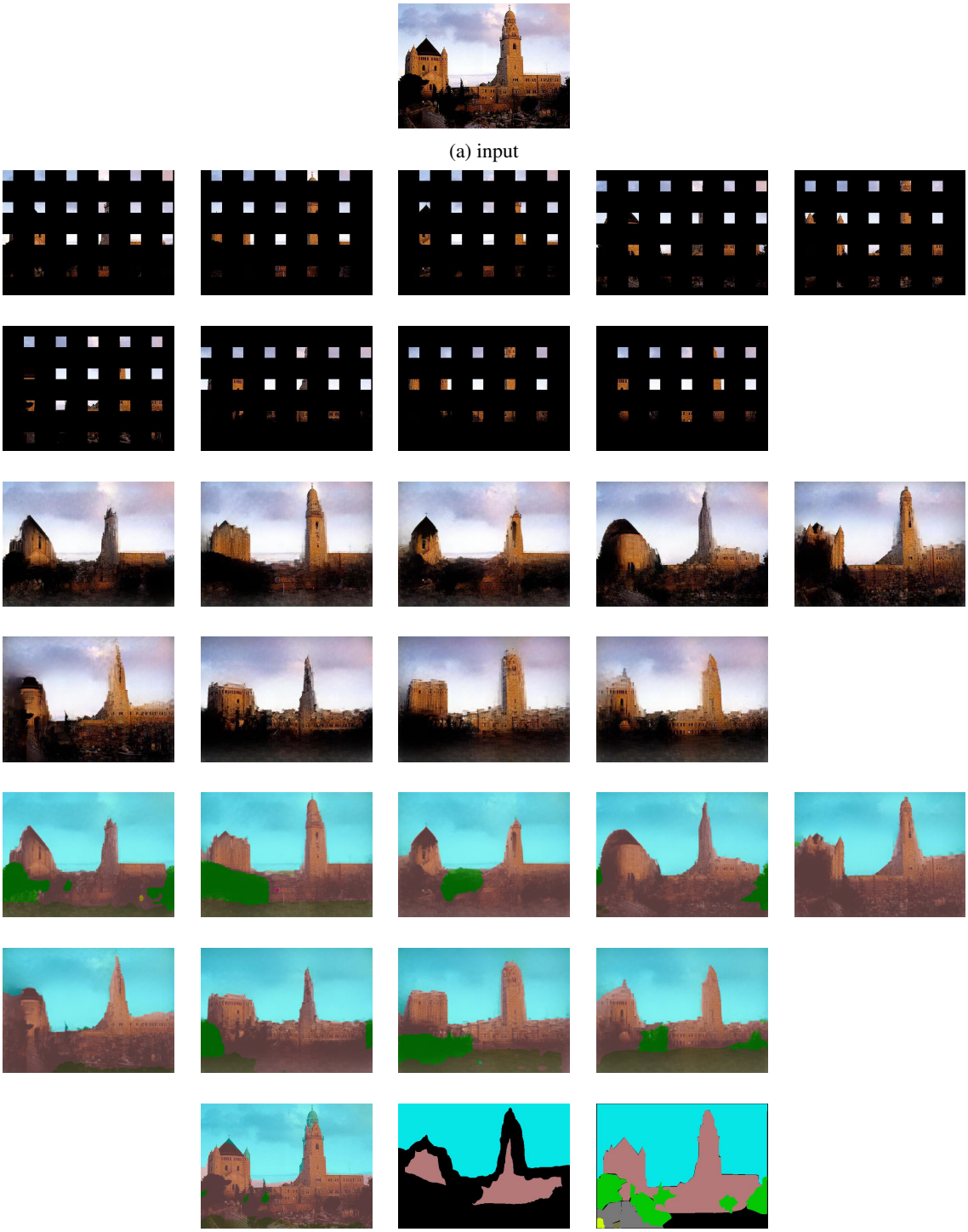
Figure 16: DEMASKED SMOOTHING recovery column masking illustration for an image from ADE20K (Zhou et al., 2017) with ZITS (Dong et al., 2022) and Senformer (Bousselham et al., 2021)



(a) input

(b) segmentation (c) certification (d) ground truth

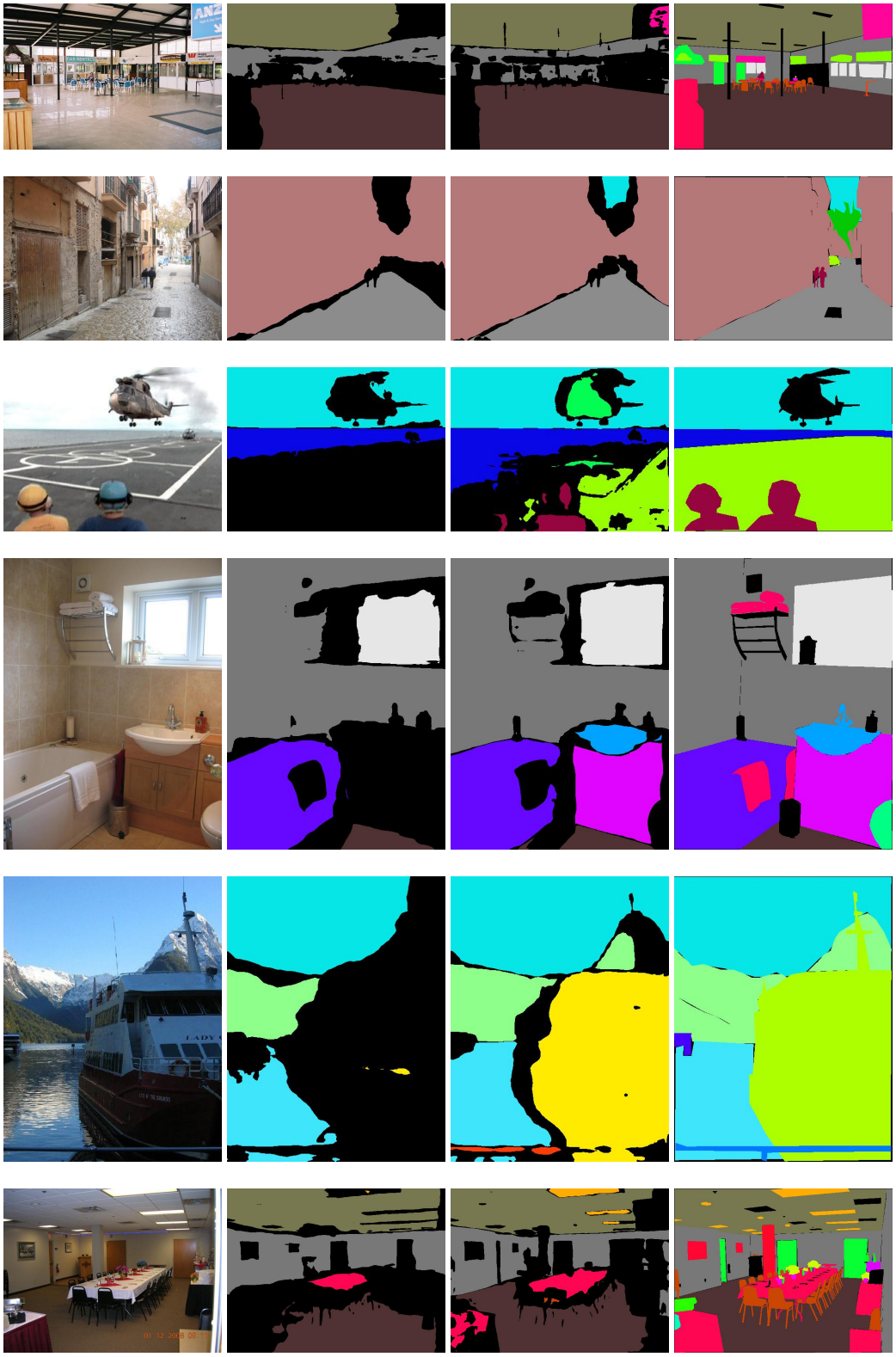
Figure 17: DEMASKED SMOOTHING recovery masking for $T = 3$, $K = 7$ masks (Section 4.1) illustration for an image from ADE20K (Zhou et al., 2017) with ZITS (Dong et al., 2022) and Senformer (Bousselham et al., 2021).



(a) input

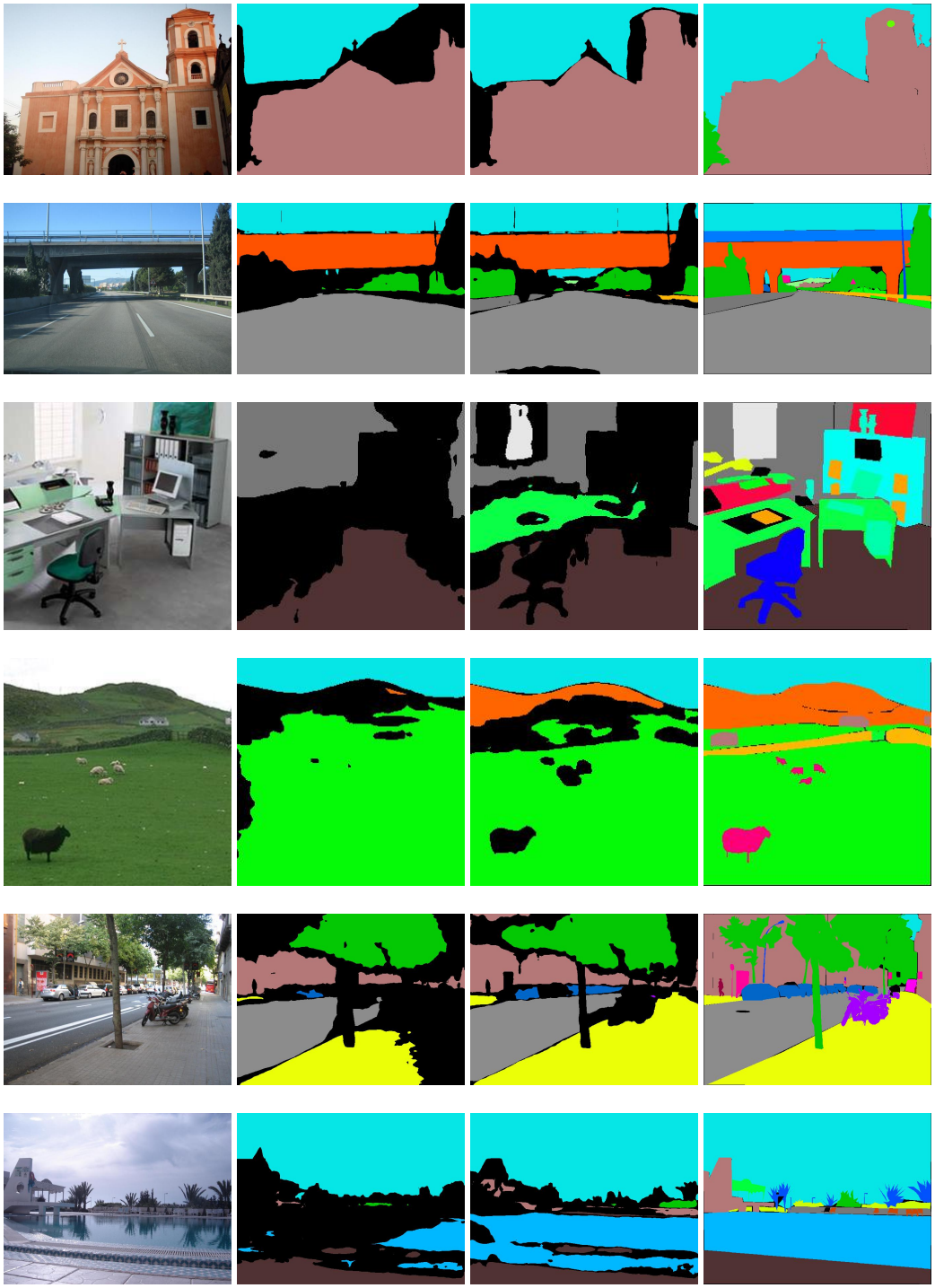
(b) segmentation (c) certification (d) ground truth

Figure 18: DEMASKED SMOOTHING recovery masking for $T = 5$, $K = 9$ masks (Section 4.1) illustration for an image from ADE20K (Zhou et al., 2017) with ZITS (Dong et al., 2022) and Senformer (Bousselham et al., 2021).



(a) original image (b) recovery map (c) detection map (d) ground truth

Figure 19: Certification map examples on ADE20K (Zhou et al., 2017) with ZITS (Dong et al., 2022) and Swin (Liu et al., 2021).



(a) original image (b) recovery map (c) detection map (d) ground truth

Figure 20: Certification map examples on ADE20K (Zhou et al., 2017) with ZITS (Dong et al., 2022) and Swin (Liu et al., 2021).