SMOOTH-REDUCE: LEVERAGING PATCHES FOR IMPROVED CERTIFIED ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Randomized smoothing (RS) has been shown to be a fast, scalable technique for certifying the robustness of deep neural network classifiers. However, methods based on RS require augmenting data with large amounts of noise, which leads to significant drops in accuracy. We propose a training-free, modified smoothing approach, Smooth-Reduce, that leverages patching and aggregation to provide improved classifier certificates. Our algorithm classifies overlapping patches extracted from an input image, and aggregates the predicted logits to certify a larger radius around the input. We study two aggregation schemes — max and mean — and show that both approaches provide better certificates in terms of certified accuracy, average certified radii and abstention rates as compared to concurrent approaches. We also provide theoretical guarantees for such certificates, and empirically show significant improvements over other randomized smoothing methods that require expensive retraining. Further, we extend our approach to videos and provide meaningful certificates for video classifiers.

1 INTRODUCTION

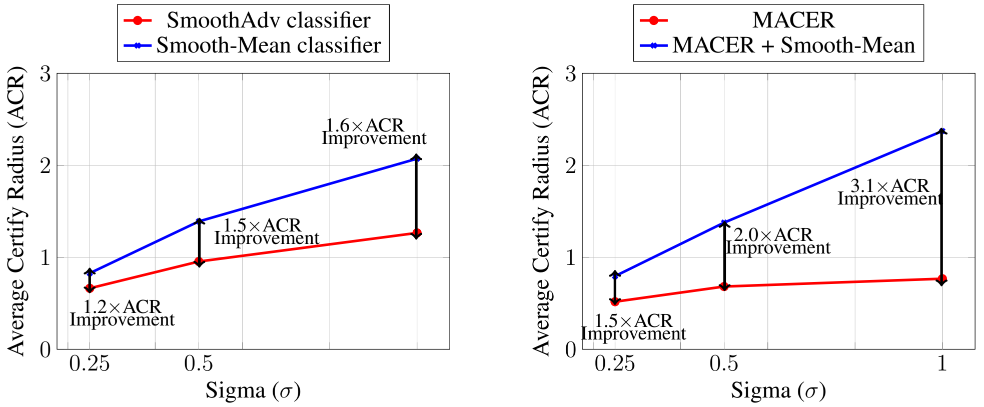
Motivation. Deep networks have been shown to be notoriously prone to "attacks" if an adversary were allowed to modify their input (Goodfellow, 2018; Szegedy et al., 2014; Athalye et al., 2018). While several heuristic "defenses" for such attacks have been proposed (Madry et al., 2018; Zhang et al., 2019; Wang et al., 2019a), only a handful of them are *provably accurate* (Wong & Kolter, 2018; Cohen et al., 2019; Salman et al., 2019), i.e., they provide guarantees for robust performance. Such defenses provide *certificates* that a deep classifier does not change its predictions within a volume (parametrized by a radius, R).

Wong & Kolter (2018) pioneered the use of bound propagation to derive upper bounds on the certification radius for networks with ReLU activations; however, this approach fails to scale to larger networks, and the bounds become vacuous quite quickly. Subsequently, Cohen et al. (2019) and Salman et al. (2019) employed *randomized smoothing* (RS) to establish bounds on the (local) Lipschitz constant of a smoothed deep classifier. Such approaches have since been fruitfully developed to provide non-vacuous certificates.

Randomized smoothing (RS) methods typically involve *convolving* the deep classifier under consideration, f, with any smooth, continuous probability distribution, $\mathcal{P}$ and deriving a radius of certification R for all points within some ℓ_p-ball around $\mathbf{x}$. For simplicity, consider certifying models in terms of ℓ_2-bounded input perturbations. Then, a randomized smoothing scheme produces a (bound on) a certificate parameter R such that $\mathbb{P}\left(\mathcal{P} \star f(\mathbf{x}) \neq \mathcal{P} \star f(\mathbf{x} + \delta)\right) \approx 0$ for any $\|\delta\|_2 \leq R$.

In practice, this functional convolution is achieved by randomly sampling noise vectors $\mathbf{z}_i \sim \mathcal{P}$, adding them to copies of the input $\mathbf{x}$, and performing inference over each copy. The resultant smooth classifier estimates the empirical probability mass, p_A for the correct class, A. Yang et al. (2020) derive the radius of certification using $R = \int_{1-p_A}^{1/2} \frac{1}{\Phi(p_A)} dp_A$, where p_A is the probability of the correct class under the noisy inference, and $\Phi(\cdot)$ is the appropriate CDF. The geometry of the ℓ_p ball influences the choice of noise; For example, Gaussian noise provides ℓ_2 certificates.

The RS approach provided the first theoretically supported non-trivial certificates for deep neural network classifiers. Nevertheless, there still remain several real-world shortcomings. First, in order to estimate empirical probabilities, one requires a large number of samples. Second, the certified accuracy achieved via randomized smoothing (RS) is substantially lower than empirical robust accuracy such as adversarial training. Third, the addition of noise to the inputs significantly degrades the performance of the network. The last problem is particularly challenging, requiring care to handle.



(a) SmoothAdv (Salman et al., 2019) vs. Smooth-Mean (b) MACER (Zhai et al., 2020) vs. Smooth-Mean

Figure 1: **Smooth-Reduce improves upon SmoothAdv and enhances MACER**. Smooth-Reduce leverages patching to emulate ensembles to reduce variance of smooth predictions by the base classifier. We study two flavors of Smooth-Reduce that use *max* and *mean* aggregation schemes respectively. Smooth-Reduce takes any base classifier that is trained to be robust to noise and uses a RS-inspired certification algorithm to derive larger certificates with lower variance. Our approach shows significant improvements over concurrent smoothing methods in certified accuracy and abstention rates across several datasets and classifiers.

Typical RS methods (such as Salman et al. (2019) and Cohen et al. (2019)) propose noise-augmented training to sidestep this problem. However, in practice we see that noise-augmented training comes with a price: the certified accuracies drop off significantly as the radius increases. Several other approaches (Zhai et al., 2020; Alfarra et al., 2020; Horv'ath et al., 2021) propose improvements to prevent such a dramatic drop-off, but they involve careful model re-training with noise augmentation, often involving several heuristic parameters.

How then can we get better certificates? To start, observe that any RS scheme involves two basic components: the base classifier f and the noise distribution $\mathcal{P}$. Adding noise provides certified robustness, but decreases accuracy; resolving this tradeoff is the key. Works such as Alfarra et al. (2020); Súkeník et al. (2021); Yang et al. (2020) focus on $\mathcal{P}$, and propose convolution with more sophisticated (sometimes even data-dependent) noise distributions. On the flip side, works such as Salman et al. (2019); Zhai et al. (2020); Addepalli et al. (2021); Horv'ath et al. (2021); Wang (2021); Jeong et al. (2021) focus on training better base classifiers f. We pursue the latter approach in this paper.

At the heart of our approach is a simple technique that is ubiquitous in machine learning inference: *ensembling*. Aggregating results from an ensemble of *diverse* classifiers acting on a given data point has long been used to improve classifier performance in standard (non-adversarial) inference settings. However, in practice, training large (and diverse) ensembles for deep networks can be non-trivial (and sometimes even prohibitively expensive). The difficulty compounds when noise augmentation and adversarial training are considered.

We overcome this difficulty by *emulating* an ensemble classifier by extracting a set of (large) patches from a given image, running a (single) base classifier on all these patches, and aggregating the results. This technique has also been successfully employed by a recent series of patch-level models (Dosovitskiy et al., 2020; Trockman & Kolter, 2022). Specifically, we posit that small affine transformations of an image induce sufficient diversity leading to more robust performance. Further, we also study two popular aggregation schemes for ensembling — *max* and *mean* aggregation — and demonstrate that both significantly outperform all existing RS approaches.

Contributions. We propose an adaptive, ensembling-based *training-free* smoothed classifier that significantly outperforms existing RS methods.

Our specific contributions are as follows:

1. We present a modified smooth classifier that leverages an input set constructed by extracting patches of the input image, and achieves a higher certified radius using aggregation.
2. We show that our certificates hold with high probability with intuitive extensions of the theoretical analysis by Salman *et al.* (Salman et al., 2019).
3. We demonstrate significant improvements in certification performance for CIFAR-10 and ImageNet compared with several state-of-the-art randomized smoothing approaches.

4. Finally, we extend our approach to provide certificates for video classifiers on UCF-101, therefore demonstrating that our approach scales to high dimensional domains.

Techniques. Our approach consists of four steps: (1) We emulate a diverse set of inputs from a given (single) image or video input. For images, we sample overlapping contiguous patches. For videos, we sample overlapping subvideos from the video stream. (2) We then follow the standard randomized smoothing approach by creating n copies of the input set, and adding independent noise vectors to each element in all copies (3) We then estimate the predicted probability of each class for each copy of the input set, and average aggregate of the estimated probabilites. (4) Finally, we record certificates for the input using Corollary 1 below. We explain each of these steps in detail in Section 2.

Certified Defenses. Ever since deep classifiers have been found to be vulnerable to adversarial attacks (Szegedy et al., 2014; Athalye et al., 2018; Carlini & Wagner, 2017), considerable efforts have been directed towards developing reliable defenses (Madry et al., 2018; Zhang et al., 2019; Wang et al., 2019b; Samangouei et al., 2018; Yin et al., 2020). The above approaches lack strong theoretical guarantees. Provable defenses (that provide certificates of correctness) fall into two major categories. The first category involves establishing upper bounds on the perturbation radii for the inputs of each layer (using linear, quadratic, convex, or even mixed-integer programming) and propagating these bounds to achieve a certificate for an entire network. These include works such as Wong & Kolter (2018); Raghunathan et al. (2018a;b); Tjeng et al. (2017); Katz et al. (2017a;b); Carlini et al. (2017); Huang et al. (2017); Weng et al. (2018). However, such approaches are computationally very expensive and do not scale to large, modern deep network classifiers.

Randomized Smoothing. The second category of provable defense involves some variation of randomized smoothing (RS), which advocate "smoothing" the outputs of non-linear, non-Lipschitz networks by their functional convolution with specially-chosen noise distributions. Early works such as Lecuyer et al. (2019); Cohen et al. (2019); Salman et al. (2019) provide ℓ_2 robustness certificates by adding Gaussian noise. Subseqent works (Teng et al., 2019; Levine & Feizi, 2021) have presented certificates for the ℓ_1 and Wasserstein metrics respectively using Laplacian smoothing. Yang et al. (2020) provide a general approach to selecting distributions for various classes of adversarial attacks; unfortunately, certificates other than the ℓ_2-norm have $\Omega(d^{-1/2})$ dependence, leading to trivial certificates for high dimensional inputs.

Practically, all RS methods still fall short of heuristic empirical defense methods in terms of robust accuracy. Therefore, several works proposed modifications to the certification scheme to improve performance. MACER (Zhai et al., 2020) maximizes surrogates of the certified radius to train better certifiable models, while Alfarra et al. (2020) finetune the variance of noise for each input data point. Jeong et al. (2021) uses an adversarial version of MixUp (Zhang et al., 2018) to train models with better tradeoffs on accuracy and certifiable robustness. Notice, however, that all these approaches involve re-training large-scale models with different objectives and data augmentations.

Ensembled Defenses. Ensembling is one of the primary motivations for our Smooth-Reduce method. Horv'ath et al. (2021) propose ensembling over diverse classifiers and show that this decreases variance of predictions, allowing better certificates.Yang et al. (2021) prove that diversified gradients and large confidence margins are necessary and sufficient conditions for robust ensembles. Liu et al. (2020) propose a weighted ensemble of networks as the base classifier and demonstrate they provide better certificates. While all these approaches rely on model-level ensembling, we emulate ensembles by using patching and basic linear operations. This allows us to ensure diversity as well as improved base classifier performance. We also demonstrate that our approach outperforms ensemble smoothing by a large margin.

2 The Smooth-Reduce Framework

Preliminaries: Let $\mathbf{x} \in \mathbb{R}^d$ be a given input. For ease of exposition, we suppose that $\mathbf{x}$ is an image (and extend the framework to video inputs later below). Let $f : \mathbb{R}^d \to [0,1]^c$ be a classifier that takes the input and assigns each class label c with probability $f(\cdot)_c$. Cohen *et al.* (Cohen et al., 2019) propose performing inference using the "smooth" classifier: $\hat{f} = \arg\max_c \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(0,\sigma^2\mathbf{I})}\left[f_c(\mathbf{x}+\mathbf{z})\right]$, which enjoys the benefits of guarantees of correctness. We estimate the (most probable) class $c_A \in [C]$ and the second most probable class $c_B \in [C]$, as predicted by $\hat{f}$. It also estimates upper and lower bounds (respectively), $\underline{p_A}, \overline{p_B}$, on the corresponding class probabilities. To do so, we create n_0 copies of the input, add n_0 i.i.d. Gaussian noise vectors sampled from $\mathcal{N}(0,\sigma^2)$ and estimate p_A. The certified radius is then derived using the relation:

$$R = \frac{\sigma}{2}\left(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})\right), \tag{1}$$

where $\Phi^{-1}(\cdot)$ is the inverse Gaussian CDF (see Cohen et al. (2019); Salman et al. (2019) for a rigorous derivation). Notice that the procedure makes no assumptions on the base classifier f, and can be used to achieve a certified radius for any model (including deep classifiers).

According to Eq. 1, in order to obtain a higher radius of certification, we can increase either the noise variance or the estimated probability of the true class. However, adding large amounts of noise to the input leads to degradation in the performance of $\hat{f}$ (compared to f), and could give poor classification performance. Indeed, the majority of works focus on training deep classifiers f that are robust to noisy inputs. Instead, we focus on obtaining an improved estimate of $\underline{p_A}$.

Smooth-Reduce One approach to obtaining high-quality predictions is via *ensembling*. Using an ensemble of classifiers tends to decrease the variance of the predicted probabilities while improving accuracy (Goodfellow et al., 2016, P. 256). However, deep networks are very expensive to train, and training a large (and diverse) set of deep classifiers for a given training dataset can be prohibitive. This challenge is exacerbated in RS approaches which tend to require re-training models with noise augmentation.

Instead, we draw inspiration from the folklore practice of using *cropping* during inference to improve performance, as well as recent empirical observations regarding the considerable effectiveness of patch-based classification (Dosovitskiy et al., 2020; Trockman & Kolter, 2022). We propose patching as a mechanism to create a diverse set of (sub)images from a single input image; this allows us to emulate an ensemble while using a single base classifier.

Our method works as follows. We create a set of inputs, $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_k\}$ from a given input (base) image, $\mathbf{x}$, by using sampling (uniformly at random) sub-images with d' total pixels (with $d' < d$) and upsampling each sub-image to the original resolution (with d pixels).We then define a set of classifiers, $\{f_i : \mathbb{R}^d \to [0,1]^c\}$ such that $f_i(\mathbf{x}) = f(\mathbf{P}_i \cdot \mathbf{x}) = f(\mathbf{x}_i)$ where $\mathbf{P}_i$ is a linear patch selection operator. All base images and patches are assumed to be square for simplicity. We then define a modified version of the smooth classifier, $\widehat{f}$ that we call the "Smooth-Reduce" classifier:

$$\bar{f}(\mathbf{x}) \quad = \arg\max_c \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(0,\sigma^2\mathbf{I})}\ \textsc{Agg}_{i=1}^{k}(f_i(\mathbf{x}+\mathbf{z})_c) \tag{2}$$

where $\textsc{Agg}_{i=1}^{k}$ is a routine that *reduces* (combines) the predicted logits for inputs enumerated over the set $\mathcal{X}$. We consider two specific aggregation functions, *max* and *mean* over the predicted logits.

The Smooth-Reduce classifier, $\bar{f}$, is a simple modification of the standard RS approach. However, we find that it improves over $\hat{f}$ in two important aspects. Firstly, since it emulates an ensemble of classifiers, the variance of the estimated probability $\underline{p_A}$ is reduced, leading to sharper bounds on $\underline{p_A}$. (For a more in-depth discussion, see also the Appendix and (Horv'ath et al., 2021)) Further, we find that it also increases the estimated probability values p_A themselves; both aggregation options in Smooth-Reduce lead to more confident classification probabilities than the base classifier f. For this to hold, we have to ensure that the patches are large enough (so that meaningful classification is achieved), and that we extract sufficiently many patches from the input image (getting boosts via aggregation).

Theoretical Analysis To derive certificates of performance for our proposed Smooth-Reduce classifier $\bar{f}$, we need to rethink Eq. 1 when used with the new class probabilities. We first restate:

Theorem 1 (taken from Cohen et al. (2019))**.** *Let $c_A, c_B \in [C]$ be the most likely and second-most likely classes, and $\underline{p_A}, \overline{p_B} \in [0,1]$ be the probability estimates associated with c_A and c_B. If $\mathbb{P}_{\mathbf{z}}\left(f(\mathbf{x}+\mathbf{z}) = c_A\right) \geq \underline{p}_A \geq \bar{p_B} \geq \max_{c\neq c_A} \mathbb{P}_{\mathbf{z}}\left(f(\mathbf{x}+\mathbf{z}) = c\right)$, then $\hat{f}(\mathbf{x}+\delta) = c_A$ for all vectors δ satisfying $\|\delta\|_2 \leq R$, where $R = \frac{\sigma_{\mathbf{z}}}{2}\left(\Phi^{-1}(\underline{p}_A) - \Phi^{-1}(\bar{p}_B)\right)$.*

The derivation for the certified radius for $\hat{f}$ does not assume anything about the base classifier. We can therefore plug in our modified base classifier, $\textsc{Agg}_{i=1}^{k} f(\mathbf{x}_i)$ and similarly prove that the radius will hold with high probability.

Corollary 1 (Smooth-Reduce certificates)**.** *Let $c_A \in [C]$, and $\underline{p_A}', \overline{p_B}' \in [0,1]$ be the probability estimates from the Smooth-Reduce classifier, $\bar{f}$. If $\mathbb{P}_{\mathbf{z}}\left(\bar{f}(\mathbf{x}+\mathbf{z}) = c_A\right) \geq \underline{p_A}' \geq \overline{p_B}' \geq \max_{c\neq c_A} \mathbb{P}_{\mathbf{z}}\left(\bar{f}(\mathbf{x}+\mathbf{z}) = c\right)$, then $\bar{f}(\mathbf{x}+\delta) = c_A$ for all δ satisfying $\|\delta\|_2 \leq R$, where*

$$R = \frac{\sigma_{\mathbf{z}}}{2}\left(\Phi^{-1}(\underline{p_A}') - \Phi^{-1}(\overline{p_B}')\right). \tag{3}$$

Algorithm 1 Smooth-Reduce Certification Algorithm

```
# certify the robustness of $\bar{f}$ around $x$
function Certify($f$, $\sigma$, $x$, $n_0$, $n$, $\alpha$)
    counts0 ←SmoothReduceUnderNoise($f, x, n_0, \sigma$)
    $\hat{c}_A$ ← top index in counts0
    counts ← SmoothReduceUnderNoise($f, x, n, \sigma$)
    $\underline{p_A}$ ← LowerConfBound(counts[$\hat{c}_A$], $n$, $1 - \alpha$)
    if $\underline{p_A} > \frac{1}{2}$ return prediction $\hat{c}_A$ and radius $\sigma\,\Phi^{-1}(\underline{p_A})$
    else return ABSTAIN

# Sampling with Smooth-Reduce classifiers
function SmoothReduceUnderNoise($f$, $x, n, \sigma$)
    counts ← $[0, 0, ...C$ times$]$
    for $j = 1 : n$
        SamplePatchOperator() → $\{P_i\}$ # Can be dense or random sampling.
        $\{z_i\}$ → Sample from $\mathcal{N}(0, \sigma^2\mathbf{I})$
        $\{\hat{y}_i\} = \{f(P_i \cdot (x + z_i))\}$
        $\hat{y}_j$ = Reduce($\{\hat{y}_i\}$) # Reduce over patches
        counts[$\arg\max_{c \in C} \hat{y}_j$]$+ = 1$
    return counts
```

Following standard practice in evaluating RS algorithms, we modify the Predict and Certify subroutines as in Cohen et al. (2019). For the prediction step, we create n copies of our input set. We then modify the routines appropriately with the aggregation step; see Alg. 1 in the appendix for the pseudocode and Fig. 6 for a diagram of the process. Note for Smooth-Max, we scale the predicted logits using softmax over the classes for each copy. Similar to Cohen et al. (2019), our classifier abstains unless the event, $\underline{p'_A} \geq 1/2$ holds with probability larger than $1 - \alpha$.

Notice that since we are estimating the lower bound on $\underline{p_A}$, the robustness guarantee holds in high probability. However, we can leverage the benefits of ensembling in each step to improve the success probability. For example, consider that there exists an adversarial example δ for any sub-classifier f_i such that $\|\delta\|_2 \leq R$. Suppose the probability of such an event occurring can be upper bounded by α. Then, the probability of δ to be an adversarial example for $\bar{f}$ is at most α/k; see the appendix for a detailed discussion. Theoretically, this allows us to achieve the same performance as $\hat{f}$ with k times fewer samples. However, in practice, this may lead to a high abstention rate if the base classifier f is itself not robust enough to noise.

Remark: A concern that arises here is if we are actually certifying the original image, as practically, we add noise to the patches and not the image. However, note that the patch classifiers can be rewritten as $f_i(\mathbf{x}) = f(\mathbf{P}_i\dot{\mathbf{x}})$, where $\mathbf{P}_i$ is an linear operator. In our approach, we add Gaussian noise, $\mathbf{z}_i$ to the patches, $f_i(\mathbf{x}_i + \mathbf{z}_i) = f(\mathbf{P}_i(\mathbf{x} + \mathbf{z}))$. We then consider $\mathbf{P}_i$ to be a part of the new Smooth-reduce classifier with all the corresponding guarantees. As long as $\mathbf{P}_i$ is linear, the standard deviation of the input noise still remains the same and Thm. 1 holds.

Smooth-Max versus Smooth-Mean. By construction, if the base classifier succeeds on patches then Smooth-Max should intuitively perform at least as well than standard randomized smoothing. During inference, the Smooth-Max classifier picks the best of possible patches in the input set. Therefore, in expectation, we hope that the Smooth-Max classifier will be more robust. The Smooth-Mean classifier, on the other hand, improves predictions using averaging to reduce the variance. Intuitively, patches that are classified with low confidence are countered by patches with very high confidence. Our observation is that the Smooth-Mean classifier abstains less frequently as compared to Smooth-Max and base RS classifiers. This also showcases one of the limitations of Smooth-Max classifiers: if the base classifier, f is very robust to noise, then a bad patch can consistently be chosen leading to the Smooth-Max classifier abstaining more often. In such a case, the Smooth-Mean classifier rectifies this by not relying on a single patch, leading to fewer abstentions. We also see evidence of this behavior in our results as discussed below.

3 Experiments and Results

Certificates for Image Classifiers We evaluate our approach by certifying classifiers trained on CIFAR-10 and ImageNet. To meaasure the performance, we consider three metrics: (1) The approximate certified accuracy with respect to the radius, (2) Average Certified Radius (ACR), and (3) the abstention rate. We define average certified radius as in Zhai et al. (2020), where for each (x_i, y_i) in the test set, D_{test}, and the corresponding certified radius, R_i, we calculate the ACR as $\frac{1}{|D_{\text{test}}|}\sum_{(x_i,y_i)} \mathbf{1}[\bar{f}(x_i) = y_i]R_i$. The abstention rate is defined as the fraction of abstentions for the given test set. We show that our approach improves upon all the metrics over other randomized smoothing methods.

Setup: We use the base classifiers with the highest reported performance trained by Salman et al. (2019) for CIFAR-10 and ImageNet. The base classifiers have been adversarially trained to be robust to varying Gaussian distributions as well as smooth adversarial attacks. Further, we use $n_0 = 100$ samples for prediction, and $n = 100k$ for certifying CIFAR-10, and similarly $n_0 = 100, n = 100k$ for Imagenet. We choose the best reported models in Zhai et al. (2020); Alfarra et al. (2020) and Horv'ath et al. (2021) for comparisons with the same setting unless otherwise stated. We also retrained models for MACER (Zhai et al., 2020) for CIFAR-10 and Imagenet. For others, we report the numbers from literature. Additional details are available in the appendix.

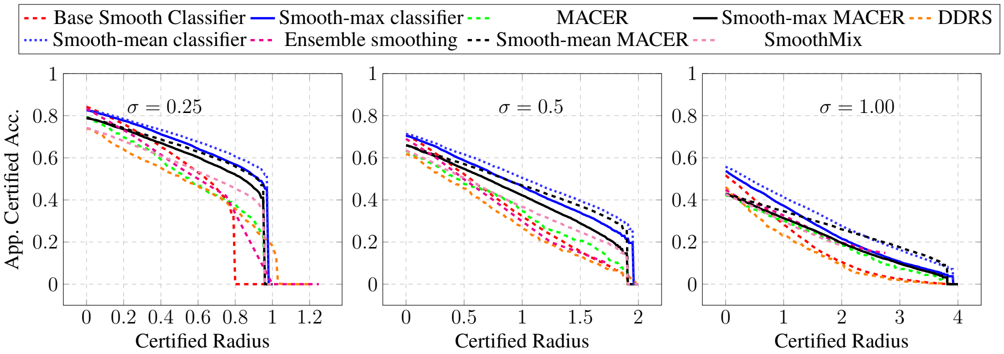


Figure 2: **Certified Accuracies for Cifar-10.***Smooth-Max classifiers provide better certificates as compared to other approaches. For the same number of copies of the input set, we see that Smooth-Reduce outperforms both in terms of certified radius and also abstains less frequently than all other approaches. Additionally, Smooth-Reduce can be effortlessly integrated with other improvements in RS (for example, MACER (Zhai et al., 2020)) without any retraining to achieve improved certificates.*

Results on CIFAR-10: For CIFAR-10, we use the pretrained Resnet-110 from (Salman et al., 2019). Since the inputs are required to be 32×32 images, we resize each input image to be 36×36 and sample 4 patches of the size 32×32 using either a random or uniform sampling process. We certify the CIFAR-10 test dataset with both variants of our Smooth-Reduce algorithm for $\sigma = 0.25, 0.5, 1.0$. Note that we use the corresponding adversarially trained noise models from (Salman et al., 2019) as our base models. Fig. 2 shows results our experiments. We also compare Smooth-Reduce with SmoothAdv (Salman et al., 2019), DDRS (Alfarra et al., 2020), MACER (Zhai et al., 2020), Ensemble smoothing (Horv'ath et al., 2021) and SmoothMix (Jeong et al., 2021). We use either use the best models shared by the authors where ever possible. Note that the Smooth-Mean and Smooth-Max algorithms outperform other approaches by a significant margin in terms of certified accuracy. Further, the average certified radius for Smooth-Mean and Smooth-Max exceed that of other approaches by at least 25% and 14% respectively. We also see that the improvements increase in magnitude as the noise variance increases. Finally, Smooth-Reduce is successful in reducing abstention rates. T. 1 shows the average certified radii and abstention rates for the various certification algorithms. Also see that Smooth-Mean classifiers tend to abstain far less often as compared to Smooth-Max classifiers.

We also study the effect of confidence calibration by ensuring both SmoothAdv and Smooth-Reduce classifiers see the same number of overall patches. To ensure fair comparison, we run SmoothAdv certification with $N = 100k$ and $\alpha = 0.001$ and Smooth-Reduce with $N = 10,000, k = 10$, and use α to be 0.01 for each sub-classifier. Fig. 4(b) shows that under the same number of samples, Smooth-Reduce is able to certify larger radii while improving the certified accuracy.

Table 1: **Results for CIFAR-10.**ACR and Abstention rates for CIFAR-10. The best performer is bolded and the second best is italicized. Results with a $*$ are from reported from references.

Algorithm	ACR ↑			Abst. Rate ↓		
σ	0.25	0.5	1.0	0.25	0.5	1.0
Smooth-Max (Ours)	*0.798*	1.29	1.803	0.021	0.058	0.125
Smooth-Mean (Ours)	**0.827**	**1.390**	*2.072*	**0.013**	**0.032**	*0.081*
Smooth-Max Macer (Ours)	0.765	1.236	1.920	0.026	0.0723	0.118
Smooth-Mean Macer (Ours)	0.795	*1.377*	**2.372**	0.022	*0.0478*	**0.077**
SmoothAdv (Salman et al., 2019)	0.663	0.954	1.265	0.0377	0.101	0.220
MACER (Zhai et al., 2020)	0.517	0.682	0.767	0.206	0.366	0.576
DDRS (Alfarra et al., 2020)	0.678	0.942	1.185	0.048	0.122	0.244
Ensemble RS* (Horv'ath et al., 2021)	0.583	0.756	0.788	-	-	-
SmoothMix (Jeong et al., 2021)	0.739	1.15	1.826	0.032	0.084	0.138

Table 2: **Results for Imagenet.** ACR and Abstention rates for ImageNet. The best performer is bolded and the second best is italicized. Results with a $*$ are reported from references.

Algorithm	ACR ↑			Abst. Rate ↓		
σ	0.25	0.5	1.0	0.25	0.5	1.0
Smooth-Max (Ours)	*0.767*	*1.453*	*2.611*	*0.008*	*0.038*	*0.108*
Smooth-Mean (Ours)	**0.786**	**1.513**	**2.931**	**0.002**	**0.024**	**0.048**
SmoothAdv (Salman et al., 2019)	0.729	1.327	2.204	0.02	0.098	0.22
MACER* (Zhai et al., 2020)	0.544	0.831	1.008	-	-	-
Ensemble RS* (Horv'ath et al., 2021)	0.545	0.868	1.108	-	-	-
SmoothMix* (Jeong et al., 2021)	-	0.846	1.047	-	-	-

Results on Imagenet: We also test our approach on 500 images from Imagenet[1] to certify a pretrained Resnet-50 model from Salman et al. (2019). We resize our inputs to 256×256 and sample 224×224 sub-patches. We use $4, 8,$ and 16 patches for our approach with $n_0 = 100$ and $n = 100k$ for certification. Smooth-Reduce improves upon SmoothAdv, as shown in Fig. 3. Specifically, Smooth-Mean performs the best, having 32.9% relatively higher average certified radius as compared to SmoothAdv. Smooth-Max performs the second-best. Also notice that Smooth-Max performance remains stable with the number of patches.

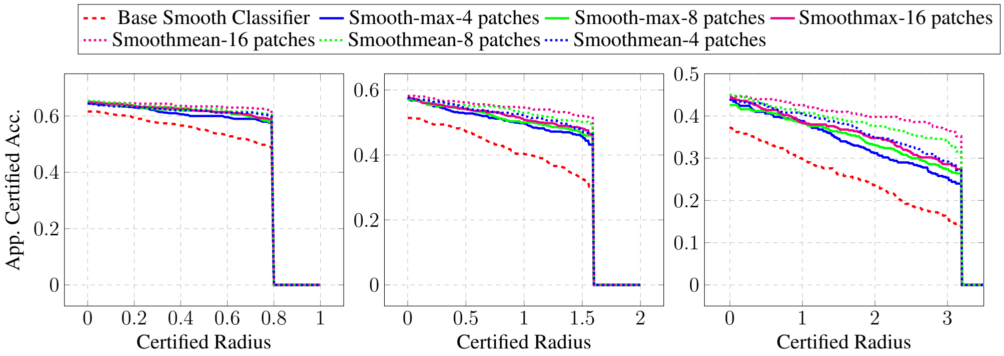


Figure 3: **Certification for ImageNet.***We see similar performance improvements of Smooth-Reduce over standard RS. Increasing number of patches does not affect Smooth-Max certificates significantly. However, Smooth-Mean classifiers have a higher approximate certified accuracy as the number of patches increase, especially as the noise variance increases. More detailed results can be found in the appendix.*

Certificates for Video Classifiers Video classifiers often employ aggregation over chunks from the video stream to tackle the problem of varying number of frames (Crasto et al., 2019)(see Fig. 9

[1] We use the subsampled Imagenet test set from Cohen et al. (2019).

in appendix). Therefore, we propose Smooth-Reduce as a natural method to certify such classifiers. While RS certificates have not been reported for such models, we observe in Fig. 5(a) and Fig. 10(see Appendix) that certified accuracies using RS are still low. As a remedy, we adapt our Smooth-Reduce algorithm to videos.

While the natural approach would be to simply look at overlapping chunks as analogues for patches, initial tests showed catastrophic loss of accuracy when we use single chunks for prediction. We therefore sample overlapping sub-videos with t frames instead. Each sub-video consists of a fixed number of chunks; each with m frames. The base video classifier aggregates over these chunks to produce a prediction. We repeat the same process of smoothing and aggregation over the sub-videos instead of chunks, and label this as Smooth-Reduce-(t, m) where t is the number of frames in each sub-video and m is the number of frames in each chunk. Fig. 9 in the appendix shows a pictorial representation;see Sec. C for a more detailed description. Note here that the base classifier itself is an ensemble over multiple 16 frame chunks.

Experiments and Results. We test our approach on 3D ResNeXt-101 RGB (Xie et al., 2017) trained on UCF-101 (Soomro et al., 2012).We retrain models initialized with weights from Crasto et al. (2019) using clips of 16 consecutive RGB frames with Gaussian noise augmentation. Similar to the setting of Crasto et al. (2019), we use SGD with weight decay of 0.0005, momentum of 0.9, and initial learning rate of 0.1. We used the first train split and the first test split for training and testing our model, respectively. Additional training details can be found in the appendix.

For inference, the video classifier follows these steps: (1) the input video stream is split into non overlapping chunks of 16 frames each, (2) the model predictions on these chunks are averaged, and returned as the output class. We run Smooth-Reduce certification by first sampling 64 frame or 128 frame sub-videos for a video stream (analogous to patching for images) to create the input set. Then we plug in the video classifier inference routine to predict classes for noisy copies of each sub-video. See Fig. 5 for results. Note that Smooth-Max and Smooth-Mean both outperform the standard randomized smoothing classifier.

Limitations of RS for Videos. We encountered several challenges while attempting to certify video classifiers. A significant challenge was training noise robust classifiers. We observe that adding Gaussian noise to video data often led to catastrophic decreases in accuracy. This could be an artifact of the architecture which averages predictions over frames by itself. Further, the memory requirements often became insurmountable to get high probability certificates. Our certificates here have been estimated using $n_0 = 10$ samples for prediction, and $n = 1000$ samples for certification with a failure probability of $\alpha = 0.001\%$. However, Smooth-Reduce allows for lower sample complexity (see appendix), allowing for Smooth-Mean models to still achieve non-trivial certified accuracy.

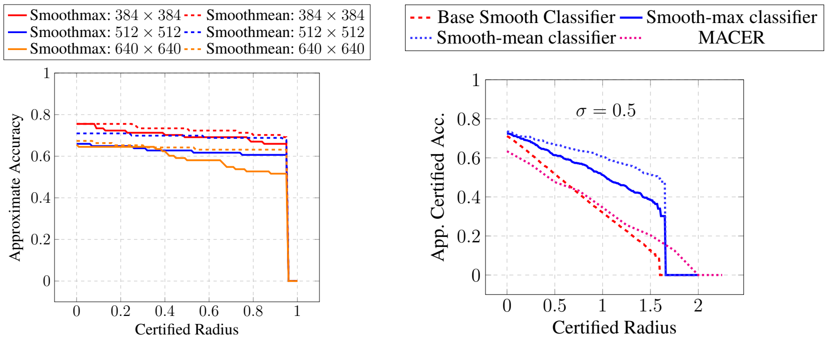


Figure 4: **Effect of resizing.(L)** *Notice that as the input size increases, the information in each patch correspondingly decreases. The base classifier performs worse overall for each patch leading to lower certified accuracy. (b)* ***Certification with Confidence Calibration (R)****. Under the same number of overall samples and calibrated failure probabilites, Smooth-Reduce out-performs SmoothAdv (Salman et al., 2019) in both certification radius and certified accuracy.*

Number of patches: We measure the effect of the number of samples used for Smooth-Reduce certification on Imagenet classifiers. Fig. 3 shows that Smooth-Max classifiers are relatively unaffected by the number of samples chosen. However, we see that increasing number of patches improves performance of Smooth-Mean certificates. This can be attributed to better empirical estimates as the

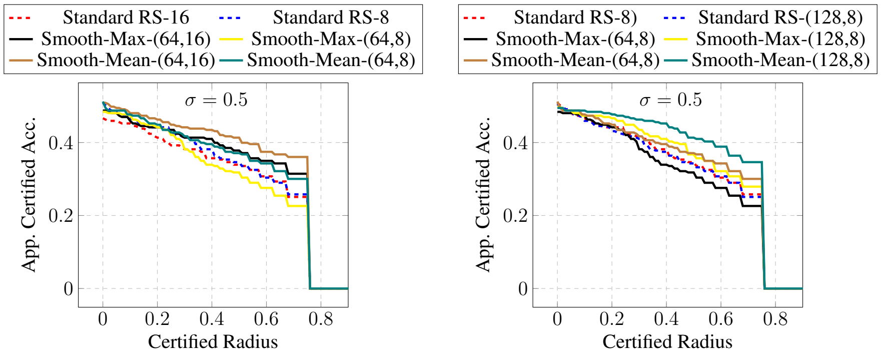


Figure 5: **Results on UCF-101.** *For SmoothAdv, the number represents the number of frames in each chunk. For Smooth-Reduce classifiers, the first number is the subvideo size, and the second is the number of frames in each chunk. (**Left**) Varying chunk sizes.Our modified Smooth-Reduce classifiers provide higher certified accuracies as compared to RS on UCF-101 videos. Also observe that for a fixed subvideo size, RS and Smooth-Reduce classifiers using larger chunks (16 frames over 8 frames) are more robust. (**Right**) Varying subvideo sizes. Smooth-Reduce classifiers using larger subvideo sizes are more robust. The difference is higher for Smooth-Mean classifiers which outperform RS by a large margin. More results can be found in the Appendix.*

number of samples increase, and also verifies our theoretical analysis; see appendix. Note too that this difference is more evident for higher noise variances.

Effect of Resizing: Another component of Smooth-Reduce is the resizing step undertaken while sampling. While theoretically it should not affect the radius, the base classifier does assume that the features would be of a certain size. We analyse the effect of the resizing step by resizing Imagenet test images to $384 \times 384, 512 \times 512$ and, 640×640 and sampling 16 patches of 224×224 randomly. In Fig. 4(a), we observe that as resizing becomes more extreme, the certified accuracy falls in tandem with base accuracy.

Random v/s dense sampling: Since sampling of patches plays a large role in creating a diverse input set, we also analyse the effect of two sampling approaches; dense and uniform random. For random sampling, we select patches randomly with replacement from the resized input image, and discarding any 'invalid' patches that fall outside the image borders. For dense sampling, we sample overlapping patches with a specified stride length. We evaluate if the sampling approach affects the certificates by sampling 25 patches for each method. We observe that the sampling process does not affect the certification process as long as the number of patches are high enough (see Fig. 7 in appendix).

Effect of subvideo/chunk sizes: We analyze the effect using different subvideo and chunk sizes on UCF101 certifcation. Specifically, we use either 64 or 128 frames for the size of subvideos, and 8 or 16 frames for the size of chunks. Fig. 5 suggest that we would benefit from having a larger subvideo or chunk sizes. In fact, we observe that using 128 frame subvideos and 64 frame chunks yield the highest certification accuracy.

4 Discussion and Conclusions

We present Smooth-Reduce, an extension of the randomized smoothing proposed in Cohen et al. (2019). We empirically and theoretically proved that Smooth-Reduce classifiers improve over randomized smoothing in terms of certified radii, and abstention rate. Our approach relies on the performance boosting properties of ensemble classifiers, which we emulate by creating an input set using patches. Our approach also does not make any assumptions on the base classifier. Therefore, Smooth-Reduce can plugged in effortlessly into other certified classifiers, such as MACER (Zhai et al., 2020).

Some limitations persist. Firstly, we require higher inference-time computation than standard RS approaches. However, note that in comparison to other ensembling approaches (Horv'ath et al., 2021; Yang et al., 2021), our method does not require training multiple classifiers. Further, we have not studied adaptive attacks for this scheme, and constructing reasonable attacks for such classifiers (and verify these certificates empirically) is a complex research question in and of itself. We leave these directions to future work.

REFERENCES

Sravanti Addepalli, Samyak Jain, Gaurang Sriramanan, and R Venkatesh Babu. Boosting adversarial robustness using feature level stochastic smoothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 93–102, 2021.

Motasem Alfarra, Adel Bibi, Philip H. S. Torr, and Bernard Ghanem. Data dependent randomized smoothing. *ArXiv*, abs/2012.04351, 2020.

A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018.

N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. *IEEE (SP)*, 2017.

N. Carlini, G. Katz, C. Barrett, and D. L. Dill. Ground-truth adversarial examples. *arXiv*, 2017.

J. Cohen, E. Rosenfeld, and Z. Kolter. Certified adversarial robustness via randomized smoothing. In *ICML*. PMLR, 2019.

Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7882–7891, 2019.

A. Dosovitskiy, L. Beyer, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

I. Goodfellow. Defense against the dark arts: An overview of adversarial example security research and future research directions. *arxiv preprint*, 1806.04169, 2018.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

Mikl'os Z. Horv'ath, Mark Niklas Müller, Marc Fischer, and Martin T. Vechev. Boosting randomized smoothing with variance reduced classifiers. *ArXiv*, abs/2106.06946, 2021.

X. Huang, M. Kwiatkowska, S. Wang, and M. Wu. Safety verification of deep neural networks. *Computer Aided Verification (CAV)*, 2017.

Jongheon Jeong, Sejun Park, Minkyu Kim, Heung-Chang Lee, Do-Guk Kim, and Jinwoo Shin. Smoothmix: Training confidence-calibrated smoothed classifiers for certified robustness. *Advances in Neural Information Processing Systems*, 34, 2021.

G. Katz, C. Barrett, D. Dill, K. Julian, and M. Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. *arXiv preprint arXiv:1702.01135*, 2017a.

G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer. Towards proving the adversarial robustness of deep neural networks. *arXiv preprint*, 2017b.

M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 656–672. IEEE, 2019.

Alexander J Levine and Soheil Feizi. Improved, deterministic smoothing for L_1 certified robustness. In *ICML*, 2021.

Chizhou Liu, Yunzhen Feng, Ranran Wang, and Bin Dong. Enhancing certified robustness of smoothed classifiers via weighted model ensembling. *ArXiv*, abs/2005.09363, 2020.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. URL `https://openreview.net/forum?id=rJzIBfZAb`.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *ICLR*, 2018a.

Aditi Raghunathan, Jacob Steinhardt, and Percy S Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *NeurIPS*, 2018b.

H. Salman, G. Yang, J. Li, P. Zhang, H. Zhang, I. Razenshteyn, and S. Bubeck. Provably robust deep learning via adversarially trained smoothed classifiers. In *NeurIPS*, 2019.

Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *ArXiv*, abs/1805.06605, 2018.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012.

Peter Súkeník, Aleksei Kuvshinov, and Stephan Günnemann. Intriguing properties of input-dependent randomized smoothing. *arXiv preprint arXiv:2110.05365*, 2021.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *International Conference on Learning Representations*, 2014.

Jiaye Teng, Guang-He Lee, and Yang Yuan. ℓ_1 adversarial robustness certificates: a randomized smoothing approach. *OpenReview*, 2019. URL `https://openreview.net/forum?id=H1lQIgrFDS`.

Vincent Tjeng, Kai Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. *arXiv preprint arXiv:1711.07356*, 2017.

Asher Trockman and J. Zico Kolter. Patches are all you need? *ArXiv*, abs/2201.09792, 2022.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*, 2019a.

Yihao Wang. Improving adversarial robustness for free with snapshot ensemble. *ArXiv*, abs/2110.03124, 2021.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*, 2019b.

Tsui-Wei Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Duane S. Boning, Inderjit S. Dhillon, and Luca Daniel. Towards fast computation of certified robustness for relu networks. In *ICML*, 2018.

E. Wong and Z. Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *ICML*. PMLR, 2018.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks, 2017.

Greg Yang, Tony Duan, Edward J. Hu, Hadi Salman, Ilya P. Razenshteyn, and Jungshian Li. Randomized smoothing of all shapes and sizes. In *ICML*, 2020.

Zhuolin Yang, Linyi Li, Xiaojun Xu, Bhavya Kailkhura, Tao Xie, and Bo Li. On the certified robustness for ensemble models and beyond. *ArXiv*, abs/2107.10873, 2021.

H. Yin, Z.and Wang, J. Wang, J. Tang, and W. Wang. Defense against adversarial attacks by low-level image transformations. *International Journal of Intelligent Systems*, 2020.

Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. Macer: Attack-free and scalable robust training via maximizing certified radius. *ArXiv*, abs/2001.02378, 2020.

H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, pp. 7472–7482, 2019.

Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412, 2018.

A Implementation

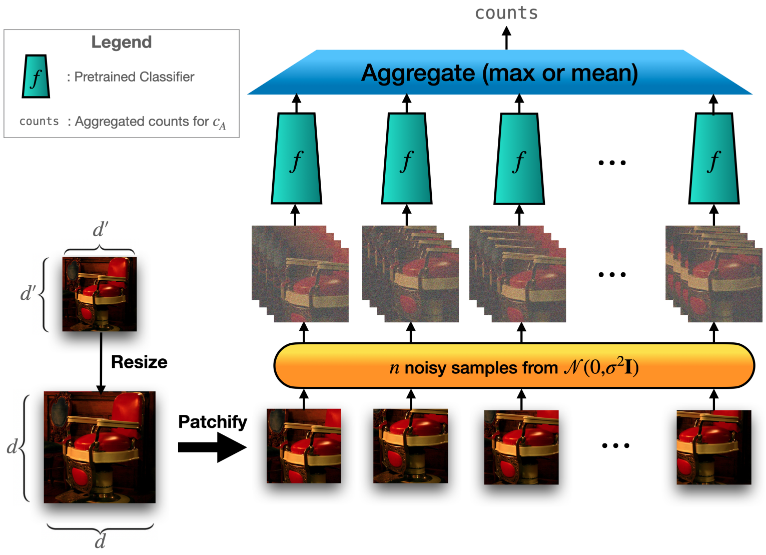

counts
Aggregate (max or mean)
Legend
f
: Pretrained Classifier
counts : Aggregated counts for c_A
f
f
f
…
f
…
d′
d′
Resize
n noisy samples from $\mathcal{N}(0,\sigma^2\mathbf{I})$
Patchify
…
d
d

Figure 6: **Smooth-Reduce Certification.***Smooth-Reduce modifies the RS certification in two ways. First, we create an input set to simulate an ensemble. We use patches sampled from the resized image. Following the* CERTIFY *subroutine from (Cohen et al., 2019), noise is added to every element in the set. Next, the counts of predicted classes are aggregated to estimate $\underline{p_A}$, the probability of the most probable class, c_A. The final step uses Eq. 3 with $\underline{p_A}$ to derive a certificate that holds with high probability.*

B Results and Discussion

We support our observations in Sec. 3 with some additional results presented here.

B.1 Additional Results for CIFAR-10

Table 3: **Detailed results on CIFAR-10.**

σ	Approach	Radii																
		0.0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0	2.25	2.5	2.75	3.0	3.25	3.5	3.75	4.0
0.25	Smooth-Max (Ours)	82.6	76.2	68.0	58.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Smooth-Mean (Ours)	82.7	77.1	70.6	62.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SmoothAdv (Salman et al., 2019)	82.3	71.7	57.9	41.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	MACER-Smooth-Max (Ours)	79.1	72.0	63.6	54.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	MACER-Smooth-Mean (Ours)	78.8	72.7	66.0	58.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	MACER (Zhai et al., 2020)	79	67	52	40	0	0	0	0	0	0	0	0	0	0	0	0	0
	DDRS (Alfarra et al., 2020)	73	61	51	39	18	0	0	0	0	0	0	0	0	0	0	0	0
	Ensemble (Horv'ath et al., 2021)	83	70	55	42	0	0	0	0	0	0	0	0	0	0	0	0	0
	SmoothMix (Jeong et al., 2021)	75.4	67.1	57.5	47.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.50	Smooth-Max (Ours)	70.5	65.5	58.7	52.4	46.4	39.9	34.0	27.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Smooth-Mean (Ours)	71.4	67.0	61.7	56.0	50.5	44.8	39.7	33.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SmoothAdv (Salman et al., 2019)	71.0	62.3	52.2	41.7	32.4	23.4	15.6	8.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	MACER-Smooth-Max (Ours)	65.9	60.7	54.8	48.6	42.2	35.6	28.9	21.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	MACER-Smooth-Mean (Ours)	66.0	61.4	57.0	52.0	46.8	42.1	37.1	31.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	MACER (Zhai et al., 2020)	63	56	47	43	34	25	20	11	0	0	0	0	0	0	0	0	0
	DDRS (Alfarra et al., 2020)	61	53	45	35	27	19	13	7	0	0	0	0	0	0	0	0	0
	Ensemble (Horv'ath et al., 2021)	65	59	49	45	38	32	26	19	0	0	0	0	0	0	0	0	0
	SmoothMix (Jeong et al., 2021)	63.5	57.2	50.5	43.4	36.7	30.6	24.7	19.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	Smooth-Max (Ours)	53.7	49.3	45.2	41.3	37.2	33.2	29.1	25.3	21.6	18.6	15.7	13.0	10.7	8.7	6.9	4.9	0.0
	Smooth-Mean (Ours)	55.7	52.3	48.5	44.7	41.3	38.0	34.7	31.2	27.7	24.6	21.5	18.8	16.3	13.7	11.3	8.8	0.0
	SmoothAdv (Salman et al., 2019)	51.6	46.2	40.0	34.1	28.4	23.0	17.9	13.9	10.5	7.5	5.2	3.7	2.5	1.6	0.8	0.1	0.0
	MACER-Smooth-Max (Ours)	42.9	39.9	37.1	34.2	31.3	28.6	25.8	22.8	19.5	16.8	14.3	12.0	9.7	7.7	5.8	3.2	0.0
	MACER-Smooth-Mean (Ours)	43.0	41.0	38.8	36.7	34.6	32.5	30.4	28.4	26.0	23.9	21.8	19.7	17.5	15.3	13.0	9.1	0.0
	MACER (Zhai et al., 2020)	42	39	35	32	30	27	25	20	18	15	12	9	7	6	4	1	0
	DDRS (Alfarra et al., 2020)	46	39	33	27	22	19	15	12	9	5	4	3	2	1	0	0	0
	Ensemble (Alfarra et al., 2020)	49	43	37	30	23	18	16	13	11	9	5	0	0	0	0	0	0

B.2 Additional Results for ImageNet

Table 4: **Detailed results on ImageNet.**

σ	Approach	Radii								
		0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
0.25	Smooth-Max-16(PGD)	65	61	0	0	0	0	0	0	0
	Smooth-Mean-16 (PGD)	64	63	0	0	0	0	0	0	0
	Smooth-Max-16 (DDN)	72	67	0	0	0	0	0	0	0
	Smooth-Mean-16 (DDN)	72	69	0	0	0	0	0	0	0
	Smooth-Max-8(PGD)	65	61	0	0	0	0	0	0	0
	Smooth-Mean-8 (PGD)	65	62	0	0	0	0	0	0	0
	Smooth-Max-8 (DDN)	72	67	0	0	0	0	0	0	0
	Smooth-Mean-8 (DDN)	72	68	0	0	0	0	0	0	0
	SmoothAdv (Salman et al., 2019)	61	55	0	0	0	0	0	0	0
0.50	Smooth-Max-16(PGD)	57	54	50	48	0	0	0	0	0
	Smooth-Mean-16 (PGD)	58	56	54	52	0	0	0	0	0
	Smooth-Max-16 (DDN)	66	61	56	51	0	0	0	0	0
	Smooth-Mean-16 (DDN)	66	64	61	59	0	0	0	0	0
	Smooth-Max-8 (PGD)	57	54	50	46	0	0	0	0	0
	Smooth-Mean-8 (PGD)	56	55	53	50	0	0	0	0	0
	Smooth-Max-8 (DDN)	66	61	56	48	0	0	0	0	0
	Smooth-Mean-8 (DDN)	65	63	59	54	0	0	0	0	0
	SmoothAdv (Salman et al., 2019)	51	47	40	32	0	0	0	0	0
1.0	Smooth-Max-16(PGD)	44	41	38	37	34	31	28	0	0
	Smooth-Mean-16 (PGD)	44	44	42	41	39	38	36	0	0
	Smooth-Max-16 (DDN)	54	50	46	42	37	33	28	0	0
	Smooth-Mean-16 (DDN)	55	53	51	50	47	43	38	0	0
	Smooth-Max-8 (PGD)	42	41	38	36	33	30	27	0	0
	Smooth-Mean-8 (PGD)	45	43	40	39	37	35	33	0	0
	Smooth-Max-8 (DDN)	54	49	44	39	35	31	25	0	0
	Smooth-Mean-8 (DDN)	54	53	50	46	42	37	34	0	0
	SmoothAdv (Salman et al., 2019)	37	33	29	26	23	18	16	0	0

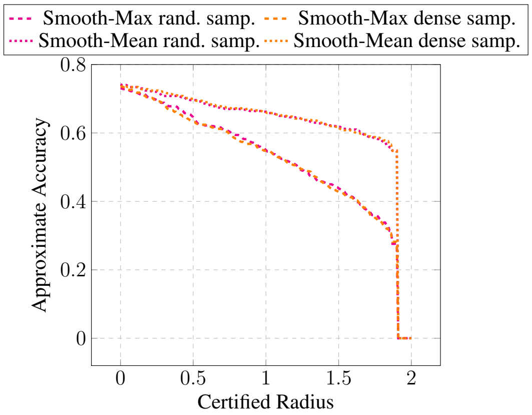


Figure 7: **Effect of Sampling Algorithm** The sampling algorithm does not affect the certified accuracy in any significant manner for both Smooth-Reduce classifiers, suggesting that the only hyperparameter of consequence is the number of patches (k).

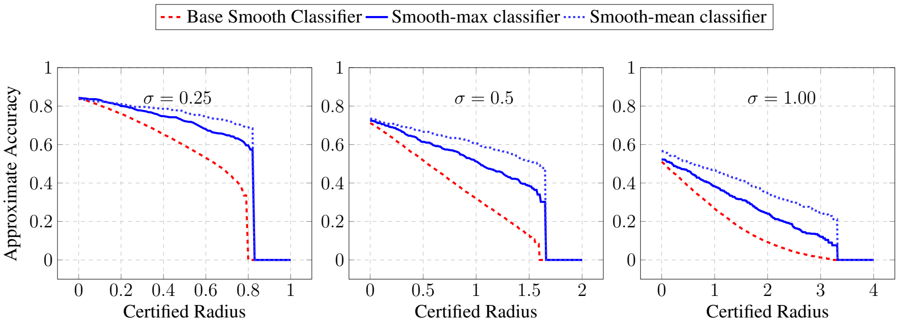


Figure 8: Confidence calibrated Smooth-Reduce

Random Sampling versus Dense Sampling. In order to understand the effect of sampling, we analyse the performance of Smooth-Reduce on CIFAR-10 under two sampling schemes: (1) randomly sampling patches under an Uniform distribution, and, (2) densely sampling patches with a specific stride length. Intuitively, the two sampling schemes should not affect performance given enough number of patches. Fig. 7 shows that this conjecture holds, with both Smooth-Max and Smooth-Mean presenting comparable performance under the two sampling schemes.

Performance with the same number of inferences. An important question that arises is if the improved certification performance is an artifact of the higher number of samples. We show that this is not the case by certifying SmoothAdv (Salman et al., 2019) and Smooth-Reduce with the same number of samples, $N = 100k$. For ensuring fair comparison, we reduce the failure rate probability rate per Smooth-Reduce sub-classifier to $\alpha = 0.01$ in comparison to $\alpha = -0.001$ for SmoothAdv. As we observe in Fig. 8, we achieve higher certified accuracies as well as better certified radii, given the same amount of compute.

C Video Classifiers

Video classifiers come in a large variety of flavors; 3d convolutional, hybrid conv-LSTM models, optical flow-based models, and others. In this paper, we only focus on certifying pure RGB frame based models. This is both due to the models being less computationally expensive as well as achieving high benign performance without a large amount of heuristic tuning. We specifically use the RGB ResNext-101 models from Crasto et al. (2019) for certifying UCF-101 videos. Crasto et al. (2019) propose a hybrid RGB-optical flow model as well, which we propose can be adapted easily to a wide variety of video classification tasks. They train two ResNext-101 models with 3D convolutions, the first on RGB frame chunks, and, the second on optical flow representations. We just use the first model for certification. However, randomised smoothing for such jointly trained multi-model classifiers is a separate and interesting technical discussion in itself.

Training. For training, we initialize our ResNext-101 with weights from the model in Crasto et al. (2019) pretrained on the Kinetics dataset. We then train the 3D CNN with 8 or 16 frame chunks from the UCF-101 training set. Following Crasto et al. (2019), we use SGD with weight decay of 0.0005, momentum of 0.9, and initial learning rate of 0.1. In order to make the classifiers robust to Gaussian noise, we also use Gaussian noise augmentation similar to Cohen et al. (2019). Further, we also use the noise-variance scheduling scheme presented in Salman et al. (2019), by slowly incrementing noise from 0 to the required noise levels every 20 epochs. For inference, the video classifier averages the logits of non-overlapping 8 or 16 frame chunks sequentially sampled from the video stream. A pictorial depiction can be seen in Fig. 9. We used the first train split and the first test split for training and testing our model, respectively. Our base model achieves $\sim 86\%$ benign accuracy on the testset. For Smooth-Reduce prediction, we follow the procedure presented above in Sec. 3 by modifying the inference step. We first sample k overlapping sub-videos from the original test video-stream. For our experiments, we use 64 and 128 frame subvideos. Next, we create n copies for each sub-video and run the base video inference described above with 16 or 8 frame chunks for each noisy copy. The predictions are then aggregated over the copies using the selected AGGREGATE (*max*/*mean*) Smooth-Reduce methods. The algorithm then returns the class with the largest count. We show results of this in Fig. 10 for noise variances of $0.25, 0.5$, and 1.0. Notice that while the certified radii are still somewhat low, Smooth-Reduce outperforms standard Randomized smoothing, certifying not only larger radii but also providing greater certified accuracy. We also see that higher chunk sizes and sub-video sizes result in better certification performance in terms of certified accuracy.

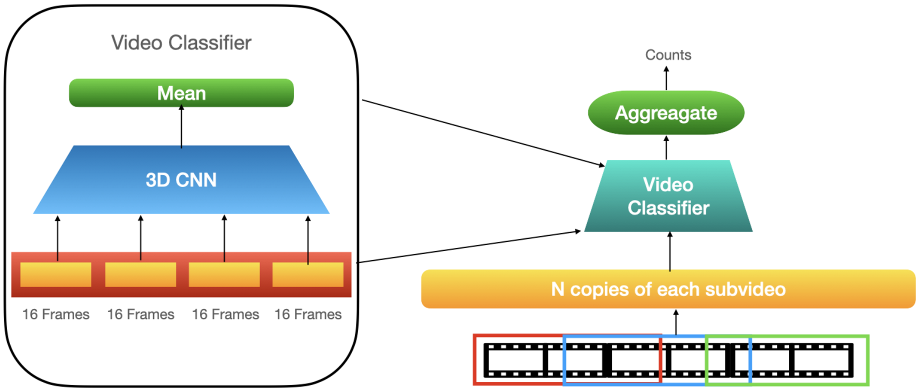


Figure 9: **Smooth-Reduce for Videos:** Video classifiers include averaging over frames or chunks of frames. Observing that larger chunk sizes provide better certificates, Smooth-Reduce takes this a step further by first sampling overlapping sub-videos with 4 or 8 chunks of 16 frames each. We then aggregate the smooth predictions over sub-videos.

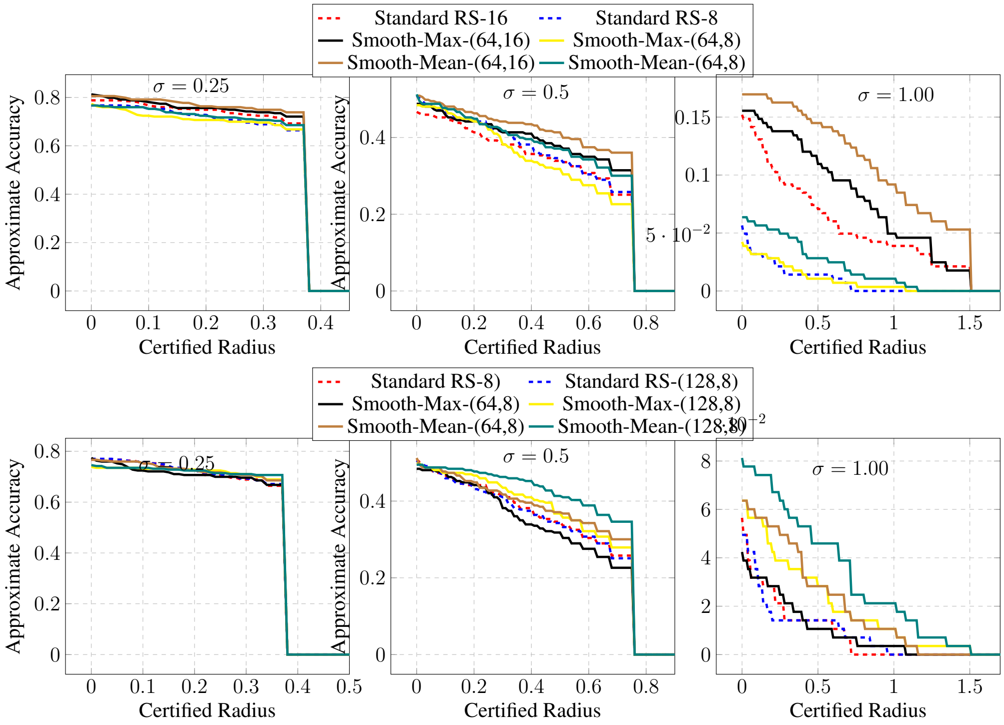


Figure 10: **Additional video results.** Notice that Smooth-Mean certifies larger radii while presenting higher certified accuracy. Another point of interest is that larger sub-videos and larger chunks show better certification performance. Model nomenclature is as follows; for standard randomized smoothing, models are named as *Standard RS*-CHUNK-SIZE; for Smooth-Reduce, we use *Smooth-*{*max/mean*}-SUB-VIDEO SIZE, CHUNK SIZE, in terms of number of frames.

D DEFERRED THEOREMS AND PROOFS

We prove that Smooth-Reduce classifiers have a lower failure probability for a given perturbation δ.

Theorem 2 (Smooth-Reduce confidence bounds). *Let $\hat{f}$ and $\bar{f}$ be the smooth and Smooth-Reduce classifiers defined above. Let f_i be the sub-classifiers in $\bar{f}$. Let R be the certified radius for $\bar{f}$ w derived using the Smooth-Reduce* CERTIFY *subroutine with n samples and k patches, with probability α_1. Let $R_i,\ i = 1 : k$ be the same for the sub-classifiers, f_i, derived using standard smoothing certification with n samples with probability α. Then, for Smooth-Mean classifiers, $\alpha_1 \leq e^{-k\alpha}(2e\alpha)^{k/2}$*

Proof. Assume that our Smooth-Mean classifier, $\bar{f}$ CERTIFY method returns some certified radius, R with the correct class, A for the given number of samples, N and patches, p. Further, we can use CERTIFY from (Cohen et al., 2019) to estimate certified radii, R_i, for each of the subclassifiers, f_i in $\bar{f}$. We assume here that the hard-classifier ensemble and the soft ensemble (that Smooth-Mean uses) are equivalent. Under this assumption, as Smooth-Mean relies on majority vote, in order for $\mathbf{x} + \delta$ to be an adversarial example, we need at least half of the classifiers to fail. To analyse this, let m_i be a Bernoulli random variable such that it takes the value 1 if classifier f_i fails and 0 otherwise. Thus,

$$\mathbb{P}[\|\delta\| < R_i] = \mathbb{P}[m_i = 1] = \alpha$$

Therefore, for $\mathbf{x} + \delta$ to be an adversarial example,

$$\mathbb{P}[\|\delta\| < R] = \mathbb{P}[\sum_{i=1}^{k} m_i \geq k/2]$$

Using a Chernoff bound (Vershynin, 2018, Thm. 2.3.1) for the sum of independent Bernoulli random variables, we get;

$$\mathbb{P}[\sum_{i=1}^{k} m_i \geq k/2] \leq e^{-k\alpha}(2e\alpha)^{k/2}$$

Note that this function decays very quickly with k, and therefore can be easily tuned to get better confidence bounds. □

While our approach relies on analysing a specific version of the adversarial example which attacks all classifiers simultaneously, we recognize that this might not be the case in general. For example, another attack may presume to make the classifier abstain every time. We do not analyse this case here, and leave the details to future work.

D.1 ANALYSING LOGITS UNDER SMOOTH-REDUCE ENSEMBLING

We further validate our claims regarding confidence intervals of Smooth-Reduce certificates by analysing the logit distribution for standard RS and Smooth-Reduce classifiers.

Setup: We study the distributions of logits for the most probable and the second most probable class for standard RS and Smooth-Reduce classifiers. For this, we consider a few test datapoints for both images, and videos and certify the best SmoothAdv classifier. Further, we certify both Smooth-Max and Smooth-Mean classifiers under the same setup. We then plot histograms of the distributions of logits. Fig. 11 and Fig. 12 show exemplars of generated histograms.

Observations and Inferences. Notice that the certified radius, R from Eq. 1 is proportional to the difference in the estimated probabilities of the two most probable classes. This difference is also proportional to the classifier margin. Therefore, in order to get better certificates, we need to ensure that the smooth-classifier presents large margins, as well higher probability estimate for the true class, c_A. Also, in order to reduce abstentions, $\underline{p_A} > 1/2$ and its variance must be low.

We see that Smooth-Max and Smooth-Mean outperform SmoothAdv on both these criterion in Fig. 11. Notice here that R depends on the difference between the means of the distribution for the most probable class (blue) and the second most probable class (orange). We see that while Smooth-Max outperforms SmoothAdv in terms of the overall proabability estimate, the margin itself is not improved much. This may lead to higher abstention rates as well as lower certificates. However, Smooth-Mean showcases not only higher estimates of $\underline{p_A}$ but also a lower variance, thus improving upon both the certified radius and probability of abstention.

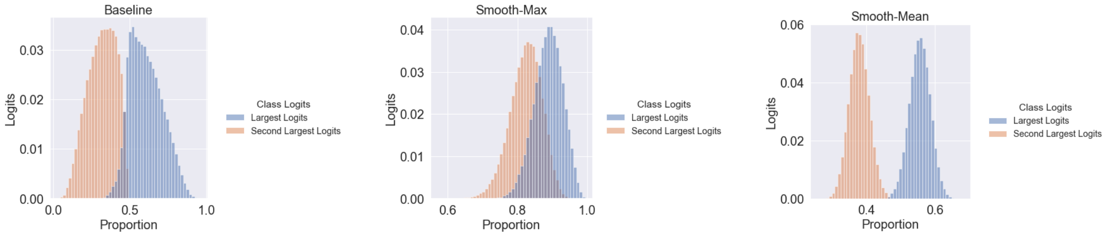


Figure 11: **Logit Distributions for Smooth Classifiers for CIFAR-10.***The histograms are arranged as follows: (L) SmoothAdv classifier, (M) Smooth-Max classifier, and (R) Smooth-Mean classifier. The blue bars the logits for the most probable class, while the orange represent those for the second-most probable class. For good smooth classifiers, the blue peak should be at* 1.0 *with low variance and the orange peak should be close to* 0*. Observe that Smooth-Max is performs better than SmoothAdv on the first criterion, while Smooth-Mean performs better on both.*

For an exemplar certificate in the case of video classifiers, we immediately observe similar behavior. In Fig. 12, we observe logit distributions for two examples from the UCF-101 dataset. It is clearly evident that while Smooth-Max and Smooth-Mean provide better margins and lower variance than standard randomized smoothing. However, the logit values are still skewed lower than those for images, and the variance across the logit values is fairly higher. This explains why our video certificates are far lower than image certificates. We conjecture that this is an effect of the difficulty in training noise-robust 3D CNN models for video. However, we leave exploring this phenomenon to future work.

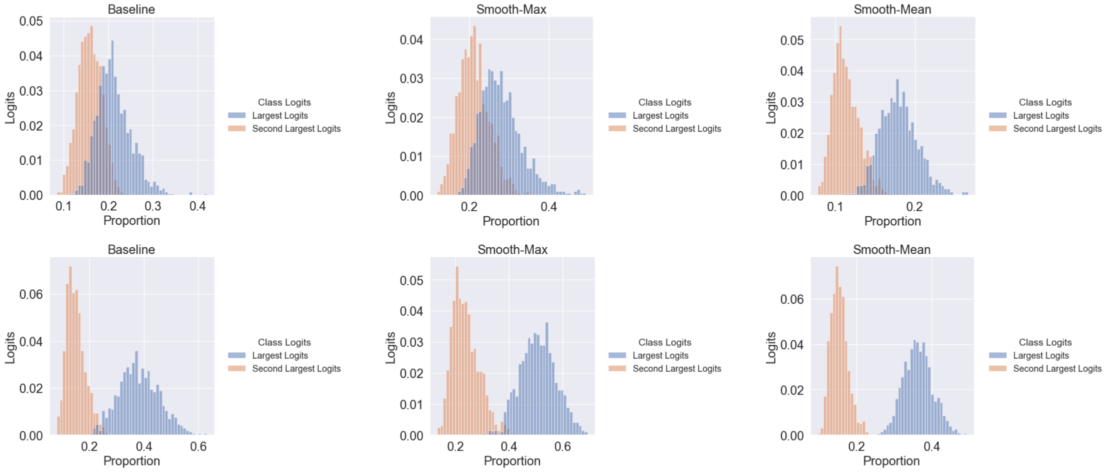


Figure 12: **Logit Distributions for Smooth Classifiers for UCF-101.** *(L) shows logit distributions for standard randomized smoothing, (M) and (R) show the same for Smooth-Max and Smooth-Mean respectively. Observe that in comparison with image classifiers, RS and Smooth-Reduce with video classifiers leads to lower* $\underline{p_A}$ *estimates as well as high variance. This results in lower certified radii and higher abstention rates as seen in Fig. 5. However, Smooth-Mean still outperforms SmoothAdv.*

D.2 Some Additional Discussion on Confidence Intervals for Ensembling

We also reproduce some results by (Horv'ath et al., 2021) which support increasing success rates for Smooth-Mean.

Horvath *et al.* (Horv'ath et al., 2021) analyse the following soft ensemble classifier,

$$\bar{f}(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{k} f_i(\mathbf{x}).$$

Let $\mathbf{y}_i$ be the logits from each of the sub-classifiers, f_i, and $\mathbf{y}$ be the same for the ensembled classifier. They further model $\mathbf{y}_i = \mathbf{y}_{i,c} + \mathbf{y}_{i,p}$, where $\mathbf{y}_{i,c}$ is a random variable representing the contribution of the i^{th} sub-classifier and $\mathbf{y}_{i,p}$ represents the contribution due to random noise added during randomized smoothing. Further they assume, $\mathbb{E}[\mathbf{y}_{i,c}] = \mathbf{c}$ and $\mathbb{E}[\mathbf{y}_{i,p}] = 0$. The variance of $\mathbf{y}_p$ is assumed to be Σ_p where $\Sigma_{ii} = \sigma_i^2$, and $\Sigma_{ij} = \sigma_i \sigma_j \rho ij$. This holds as the two processes of training and smoothing are independent. Notice that Smooth-Mean classifiers are a special class of such classifiers, where the sub-classifiers are constructed with independent sampling matrices.

Now, they analyse the class margins, $t_i = y_1 - y_1$ where y_i are elements of $\mathbf{y}$ and 1 is the majority class (WLOG). Notice,

$$\mathbb{E}[\mathbf{z}_i] = c_1 - c_i$$

$$\text{Var}[\mathbf{t}_i] = \sigma_{p,1}^2 + \sigma_{p,i}^2 + \sigma_{c,1}^2 + \sigma_{c,i}^2 - 2\sigma_{p,1}\sigma_{p,i}\rho p1, i - 2\rho c, 1i\sigma + c, 1\sigma c, i$$

Through careful arithmetic, they show that,

$$\text{Var}(\bar{\mathbf{t}}) = \sigma_p^2(k) + \sigma_c^2(k),$$

where $\sigma_m^2 = \frac{k + \binom{k}{2}\zeta_m}{k^2}(\sigma_{p,1}^2 + \sigma_{p,i}^2 - 2\rho_{p,1i}\sigma p, 1\sigma_{p,i}$ for $m \in [p, c]$, and $\zeta_m \in [0, 1]$ refers to parameter denoting covariance between $y_{i,m}$ and $y_{j,m}$; refer (Horv'ath et al., 2021) for more details.

This decoupling of the variance between the perturbations due to RS and ensembling proves to be important in understanding the benefits of ensembling. They present the following result on success probabilities,

Informal Theorem[From (Horv'ath et al., 2021)] *For a soft-ensemble of k classifiers which provides a certificate with radius R with probability $1 - \alpha_1$, the upper bound of the probability of failure decreases with $O(k^2)$*

To measure the effect on success probability, we consider the probability of a majority of the sub-classifiers predicting class 1, β_1.

$$\beta_1 = \mathbb{P}(\bar{f}(\mathbf{x} + \mathbf{z}) = 1) = \mathbb{P}(\bar{\mathbf{t}} > 0 : \forall i \in [2, C]) = \int_{\bar{\mathbf{z}} > 0 : \forall i \in [2, C]} \mathbb{P}(\bar{\mathbf{t}}).d\bar{\mathbf{z}}$$

While this integral cannot be evaluated directly as we do not know the density function for $\mathbf{z}$, we can lower bound β_1 using Chebyshev's inequality and the union bound over the incorrect $[2, C]$ classes.

$$\beta_1 \geq 1 - \sum_{i=1}^{C} \frac{(\sigma_{i,c}(k)^2 + \sigma_{i,p}(k)^2}{(c_1 - c_i)^2}$$

. As $\sigma_{i,c}$ and $\sigma_{i,k}$ decrease quadratically with increasing k, we can prove the above theorem.