Certifiably Robust Synthetic Image Detectors

Kai Zeng
University of Siena, Siena, Italy
kai.zeng@unisi.it

Mauro Barni
University of Siena, Siena, Italy
mauro.barni@unisi.it

Benedetta Tondi
University of Siena, Siena, Italy
benedetta.tondi@unisi.it

***Abstract*—Randomized Smoothing (RS) has been proposed as a method to develop deep learning classifiers with certified robustness, i.e., for which a certain level of robustness can be theoretically guaranteed. In this paper, we explore the application of the RS technique in the context of multimedia forensics, focusing on the prominent task of synthetic image detection. Our experiments, carried out on the task of detection of images generated by StyleGAN2 and Latent Diffusion models, reveal that the input pre-processing, the input size and in particular the network architecture have a noticeable impact on the certification performance. In particular, we achieved the best performance with EfficientNetB4, while we found that the certification achieved by detectors based on general-purpose features, namely CLIP, is poor. We also evaluated the performance of the RS synthetic image detectors against common image post-processing, showing that they exhibit strong robustness against a wide variety of processing, even when the distortion introduced by the processing exceeds the one the detectors can *provably* withstand.**

***Index Terms*—robust detection of AI-generated images, deepfake detection, certified robustness, adversarial machine learning**

I. Introduction

The development of forensic tools to distinguish real images from images generated by AI generative models, commonly referred to as synthetic images, is receiving increasing attention due to the importance of preserving the trustworthiness of digital media. Deep neural networks (DNNs) have been proven to be able to detect images produced by generative models with very high accuracy. However, like all DNN-based classifiers, synthetic image detectors are vulnerable to adversarial examples, i.e., imperceptible perturbation capable of misleading detectors and inducing wrong classification of fake images as real [1].

A common approach to improve the adversarial robustness of models is adversarial training [2], which consists of augmenting training with examples of attacked images. However, since adversarial training is an empirical defense, there is no guarantee that the prediction made by the robustified classifier is immune to adversarial perturbations. Indeed, robustness is often achieved only against a specific attack algorithm or a category of attacks, and such kinds of defenses can often be broken by stronger or different types of attacks [3].

When dealing with security-related applications, obtaning certified defenses that are *provably* robust to adversarial perturbations is of paramount importance. Randomized Smoothing (RS) has been proposed in machine learning research as a solution to certify the robustness of classifiers [4], [5]. An RS classifier makes a prediction on an input sample by perturbing it with Gaussian noise and deciding in favor of the class that the base classifier is most likely to predict. By doing so, a correct prediction can be theoretically guaranteed for the input within a certain radius [4]. This radius theoretically defines the *minimum* distortion an attacker must introduce to cause a misclassification, ensuring that no adversarial example can exist within this radius. Compared to other methods for robustness certification, the RS approach has gained a lot of interest due to its generality and scalability, as it can in principle be applied to *any* classifier.

In this paper, we propose to exploit the RS approach to build synthetic image detectors with certified robustness. To the best of our knowledge, this is the first time certification methods have been applied to image forensic tasks. Given that synthetic image detectors (and more in general, detectors dealing with image forensic tasks) typically rely on subtle statistical traces [6], which are more sensitive to noise, the effectiveness of the RS approach may be limited in this context, resulting in extremely small certified radii. To achieve good certification performance with the RS approach, in fact, base detectors have to be robust against (strong) Gaussian noise perturbation, which often requires training them for this purpose. Our experiments, carried out on the StyleGAN2 and Latent Diffusion (LD) detection tasks, reveal that the network architecture adopted to implement the RS detector plays a major role. In particular, a high degree of certification can be achieved using an EfficientNet-B4 base detector, while the certification is poor in the case of cutting-edge detectors based on pre-trained general-purpose features like CLIP (Contrastive Language-Image Pre-training) [7]–[9]. We also found that the input pre-processing and the image size have an impact on the performance. In particular, the normalized certified radius is larger in the case of detectors trained on images with larger input size. We also evaluate the performance of the RS detectors against common image post-processing/manipulation. Our experiments reveal that strong robustness can be achieved against a wide variety of processing operations, even when the distortion introduced by the processing exceeds the one the detector can *provably* withstand.

This work has been partially supported by the FOSTERER project, funded by the Italian Ministry within the PRIN 2022 program under contract 202289RHHP, by SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU, and by the Remedy project - New-Frontiers (PSR 2024), funded by MUR.

II. Background on Randomized Smoothing

Randomized smoothing (RS) was proposed in [4] as a model-agnostic and scalable method to develop classifiers with certified robustness. The robustness against attacks is formally guaranteed by the fact that no perturbation within a certain radius of the input -namely, the *certified radius* - can change the classification result, hence no adversarial example can be found at a distance smaller than this radius. In principle, with this method, *any* classifier can be turned into a new classifier that is certifiably robust to adversarial perturbation under L_2 norm. Specifically, the smoothed classifier is obtained from a base classifier f as follows. Given a point $x \in \mathbb{R}^d$, the smoothed classifier assigns x to the class most likely to be returned by the base classifier f when x is perturbed with addition of isotropic Gaussian noise (with standard deviation σ). In this paper, we focus on binary detectors, hence $f : \mathbb{R}^d \to \{0,1\}$. Then, the smoothed classifier g_σ is a deterministic classifier defined as:

$$g_\sigma(x) = \arg \max_{y \in \{0,1\}} P_{\mathbf{n}}[f(x+n) = y], \; \mathbf{n} \sim \mathcal{N}(0, \sigma I), \quad (1)$$

where $P_{\mathbf{n}}[A]$ denote the probability of A computed over $\mathbf{n}$. Let $\pi_0(x) := P_{\mathbf{n}}[f(x+n) = 0]$ and $\pi_1(\mathbf{x}) := 1 - \pi_0(\mathbf{x})$, and let

$$R(x, \sigma) = \sigma \Phi^{-1}(\pi_{g_\sigma(x)}(x)), \quad (2)$$

where Φ is the cumulative distribution function of the standard normal distribution $\mathcal{N}(0, I)$. It has been proven in [4] that, for any input x, g_σ classifies all the points in a L_2 neighborhood of x of radius $\delta \leq R(x, \sigma)$ in the same way. This ensures that no perturbation of x (be it adversarial or not) whose norm is lower than $R(x, \sigma)$ can modify the decision of g_σ.

In practice, the RS classifier is implemented via Monte Carlo simulations over N random i.i.d. samples $\{n_i\}_{i=1}^N$ drawn from $\mathcal{N}(0, \sigma I)$, by aggregating the decisions on noisy samples via majority voting. The aggregated average of the scores gives an approximation of the confidence $\pi_0(\mathbf{x})$ (hence $\pi_1(\mathbf{x})$). The practical classifier $\tilde{g}_\sigma$ only approximates g_σ and corresponds to g_σ when $N \to \infty$. According to [4], a very good approximation can already been achieved with $N = 10^5$. Obviously, testing the RS detector introduces an overhead, which can be regarded as the price to pay for robustness certification.

It can be observed from (2) that there is a tradeoff on the choice of the value of the noise level σ: a large σ permits to get higher certified radii R, however, it also reduces the confidence and affects the performance of g_σ. Furthermore, for the RS approach to be effective, it is necessary that the base classifier performs well under noise addition, i.e., that the samples $(x + n)$ are classified correctly by f. Therefore, in practice, it is recommended to train the base classifier f by including noise addition as augmentation. For simplicity, the same noise level is used in both the training of the base models (for data augmentation) and the testing of the RS detector.

To the best of our knowledge, no previous work has considered the application of RS to image forensic tasks, and in particular to the task of synthetic image detection, to certify the accuracy of AI-generated image detectors against attacks. Notably, multimedia forensic tasks typically rely on subtle traces that are highly sensitive to noise addition, thus challenging the application of the RS approach in this scenario.

III. Methodology

In this section, we present the specific tasks and the architectures that we considered for the base detectors. The setting and details of the implementation of RS detectors are also reported.

A. Detection Tasks and Datasets

In our experiments, we focus on the detection of fake face images generated with StyleGAN2 [10] and LD models (LDMs) [11]. For both tasks, we collected a balanced dataset of real and synthetic images. The real images are taken from the CelebA-HQ [12] (70.000) and the Flickr-Faces-HQ (FFHQ) [10] (30.000) datasets. StyleGAN2 and LD images are generated by using the pre-trained models made available by the authors in their repositories. The resolution of the images generated with these models is $1024 \times 1024 \times 3$ in the case of StyleGAN2 and $256 \times 256 \times 3$ in the case of LD. 80% of the total number of images is used for training and validation, while the remaining 20% is reserved for the tests.

B. Network Architectures

To build the base detectors, we considered two different CNN architectures, fully-trained on the synthetic image detection task, and general-purpose features from pre-trained vision language models. In particular, we considered the following:

Residual Networks (ResNet). ResNets rely on residual-based learning, which mitigates the vanishing gradient problem and enables training of very deep models [13]. In our experiments, we used ResNet-101, balancing depth and computational efficiency.

EfficientNet. In contrast with common CNNs that arbitrary scale width, depth, and input size, EfficientNet [14] employs a compound scaling strategy to optimize them We adopted EfficientNet-B4, which has demonstrated excellent performance in related tasks [15], [16]

CLIP+FCN. Recent works have shown that cutting-edge detectors can be developed using general-purpose features obtained with pre-trained vision language models and CLIP in particular [7], [8]. Following this literature we used the CLIP ViT-L/14, using the ViT-L/14 transformer architecture as image encoder. In this case, only the final fully connected layer is trained on the synthetic image detection task.

C. Settings and Metrics

1) Data processing: for the training of the base detectors, images are resized to the pre-defined input size. To evaluate the impact of the pre-processing we also consider a different pre-processing where instead of resizing the full image a random crop is first extracted and resized to the given size (using the RandomResizedCrop function from the torchvision.transforms package was used). Data standardization is performed by subtracting the mean and dividing by the standard deviation. Gaussian noise with standard deviation σ is added to the input with a probability of 0.5 as data augmentation.

2) Training settings: the model is optimized using Stochastic Gradient Descent (SGD) with a momentum of 0.9 and a weight decay of 10^{-4}. The initial learning rate is set to 0.01 for EfficientNet and 0.001 for ResNet and CLIP. A learning rate step-based decay schedule is considered, decreasing it every 30 epochs. Early stopping was employed by stopping training when the validation loss did not improve for 20 epochs.

3) Implementation of RS detectors: For every test image, $N = 10^4$ noisy versions are obtained by adding Gaussian noise with standard deviation σ. These images are then preprocessed and fed to the trained base models. The prediction of the RS detector $\hat{g}_\sigma$ at test time is then obtained as described in Section in II. [1]

4) Certification Metrics: for any test image x_i the certified radius $R(x_i, \sigma)$ is computed. Following [4], the performance of the RS detector are evaluated by measuring the *certified test accuracy* CA at a given radius R^*, namely, the fraction of test samples that are correctly classified by the RS detector and have a certified radius larger than R^*, and plotting the CA(R^*) curve. Formally, $\text{CA}(R^*) = \#\left\{x_i | \hat{g}_\sigma(x_i) = y_i \wedge R(x_i, \sigma) \geq R^*\right\} / N_t$, where N_t is the total number of test images and y_i is the ground truth label of x_i (0 = real, 1 = fake). Then, CA(0) measures the accuracy of the RS detector. We evaluated the performance on $N_t = 2000$ test images. For a given radius R^*, we can compute the corresponding Mean Square Error (MSE) of the attack that the model can withstand - i.e., the *certified* MSE - as MSE $= {R^*}^2/d$. Instead of plotting CA(R^*), we find it more convenient to plot the CA(MSE), which also makes the curves obtained for models with different input sizes comparable. As a measure of the certification capability, we report the certified MSE at selected values of CA.

IV. Results and Discussions

In this section, we report the certification performance achieved by the various trained models. We evaluated the impact of the noise, input size and network architecture on them, by focusing on the StyleGAN2 detection task. Then, some results are reported for the case of LDM, where a similar behavior is observed. Finally, the robustness of the RS detectors against common processing operations is evaluated.

A. Impact of σ, input pre-processing and resolution

TABLE I: Certification performance of RS detectors (MSE@CA), for the StyleGAN2 detection task (the resize preprocessing is applied)

RS Detector	**MSE@0.8**	**MSE@0.85**	**MSE@0.9**
ResNet 224 $\sigma = 0.12$	$3.59e-7$	$2.39e-7$	$7.97e-7$
ResNet 224 $\sigma = 0.25$	$1.06e-6$	$4.58e-7$	$3.99e-8$
ResNet 224 $\sigma = 0.5$	$2.23e-6$	$1.12e-6$	$5.98e-8$
ResNet 224 $\sigma = 1.0$	$3.03e-6$	$8.77e-7$	0
ResNet 1024 $\sigma = 1.0$	$2.44e-6$	$2.0e-6$	$1.44e-6$
ResNet 1024 $\sigma = 4.6$	$1.38e-5$	$5.96e-6$	$1.11e-6$
EfficientNet 224 $\sigma = 1.0$	$\mathbf{4.09e-5}$	$3.04e-5$	$1.28e-5$
CLIP 224 $\sigma = 1.0$	$9.37e-7$	0	0

Fig. 1 shows the CA(MSE) curves obtained for StyleGAN2, for the models trained with different values of σ, input size $224 \times 224 \times 3$ and $1024 \times 1024 \times 3$ (for $\sigma = 1$), considering the resize (a) and random-crop and-resize (b) pre-processing, when the ResNet-101 architecture is adopted. We see that, as σ increases, the accuracy of the RS detector (CA(0)) decreases, however the curves become flatter and then larger distortions can be certified.[2] This is expected as the certified radius increases linearly with σ (see (2)), even if σ also affects the second term, with a too large σ impairing the detection performance of the RS detector. By comparing Fig. 1a and Fig. 1b we also see that the input pre-processing has some impact on the certification performance, and results are better in the resize case, with larger CA achieved for similar MSE values Notably, higher CA can be obtained with larger input sizes, resulting in larger MSE@CA values. In particular, in the case of large input size, the RS detector with $\sigma = 1$ gets almost perfect accuracy (CA(0) ≈ 1). Hence, we increased the noise level to $\sigma = 4.6$ to improve the certification performance at the cost of a reduction of accuracy, exploiting the fact that higher-resolution images can tolerate stronger noise before class-distinguishing content gets destroyed.[3] Table I, rows 2-7, reports the values of the certified MSE when CA is set to 0.80, 0.85 and 0.90 for the curves in Fig. 1a.

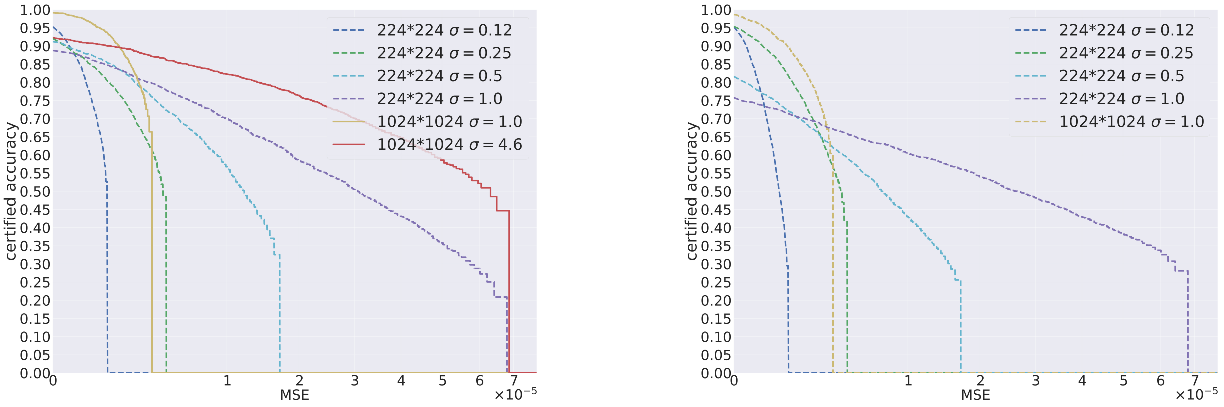
B. Impact of Model Architecture

In this section we compare the certification performance achieved with different architectures. Fig. 3 shows the results in the case of input size $224 \times 224 \times 3$ and $\sigma = 1.0$ (the comparison is similar in different settings). The resize is considered as pre-processing. We observe that the model architecture has a significant impact on the performance, with EfficientNet achieving the best results. We also see that the RS detector based on CLIP has a pretty low detection accuracy (CA(0) = 0.83). A possible reason for the poor performance of the CLIP-based RS detector is that, in the CLIP case, the embeddings are frozen and only the parameters of the fully connected layer are updated during training. However, arguably, CLIP extractor has been trained without taking into account the presence of (strong) Gaussian noise in the input. We also notice that the CA curve decreases slowly as the MSE increases in the CLIP case with respect to the case of ResNet, and there is a crossing point at MSE = 10^{-5}, where CA is 0.7. Beyond this crossing point, the CA achieved with CLIP is higher.

[1] Based on preliminary experiments we found that this number of samples is sufficient in practice and results are similar using a larger N.

[2] The presence of the cut-off value of the MSE is a consequence of the Monte Carlo simulations and of the fact that the support of the empirical distribution of the noise samples is bounded [4].

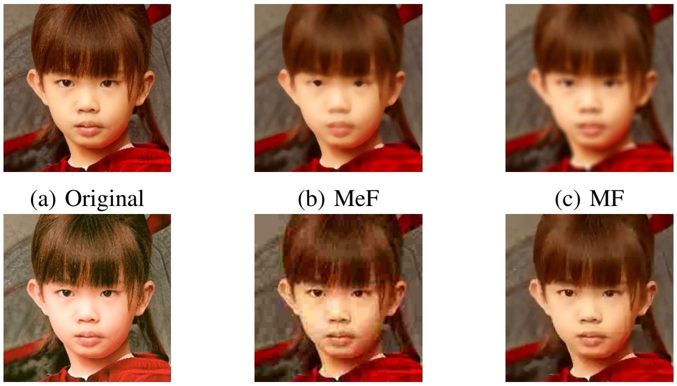
[3] The value $\sigma = 4.6$ is chosen because, for large size images (1024×1024), yields a similar $\sigma/\sqrt{d}$ and hence a similar cut-off (see [4]).



(a) Images Processed by Resize

(b) Images Processed by Random Crop and Resize

Fig. 1: Certified Accuracy curves for the various RS detectors (StyleGAN2 task). The model architecture is ResNet101.



(a) Original (b) MeF (c) MF

(d) P&S (e) JPEG QF=10 (f) WebP QF=10

Fig. 2: Examples of post-processed images.

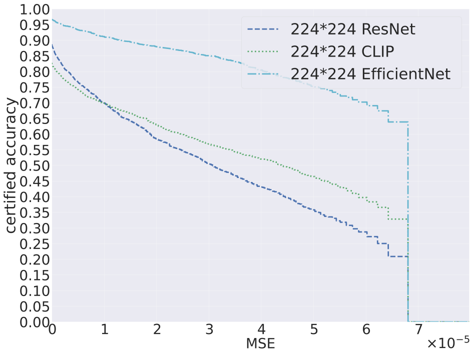


Fig. 3: Impact of model architecture on the certified accuracy (StyleGAN2 detection). The setting is the following: input size is 224×224, $\sigma = 1.0$. The pre-processing is the resize.

The values of the certified MSE at selected CA achieved with the three models are reported in Table I (lines 5, 8 and 9). Notably, when an EfficientNet is used to implement the RS detector, an MSE on the order of $10^{-4}/10^{-5}$ can be certified with a target CA of 0.80.

C. Performance on LD detection

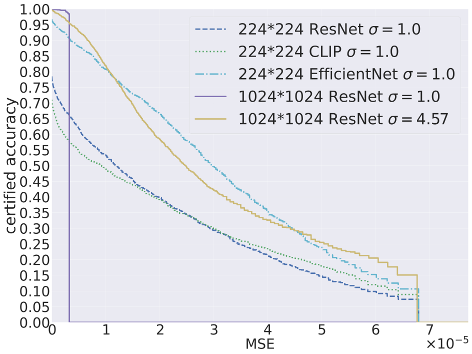


Fig. 4: Certified Accuracy curves in the case of LD detection, with different architectures and input sizes.

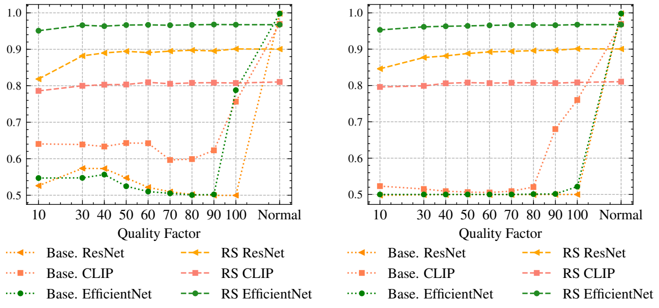
TABLE II: Accuracy of Baseline (Base.) and RS detector (RS) on post-processed images.

Process	ResNet		CLIP		EfficientNet	
	Base.	*RS*	*Base.*	*RS*	*Base.*	*RS*
MeF	0.50	**0.84**	0.78	**0.79**	0.79	**0.94**
MF	0.50	**0.75**	0.81	0.78	0.86	**0.89**
P&S	0.64	**0.84**	0.89	0.80	0.74	**0.96**

Fig. 4 shows the results we got for the RS detector trained on the LD detection task, where we can observe a similar behavior concerning the impact of the architecture and the input resolution. The resize pre-processing is considered in all cases. Specifically, by looking at the results with small size images, we see that, also in this case, much better certified robustness can be achieved using EfficientNetB4 as base model to build the RS detector, even if the MSE that can be certified is lower with respect to the StyleGAN2 case. In particular, with EfficientNet we got MSE@0.8 = 1.05×10^{-5}. As observed before, also in this case, increasing the input size helps and a larger MSE can be certified. We see that for ResNet the MSE@0.8 passes from 0 in the setting 224×224, $\sigma = 1$ to 1.07×10^{-5} in the setting 1024×1024, $\sigma = 4.6$.

D. Robustness against post-processing

We also evaluated the robustness of the RS synthetic image detectors to post-processing operations, considering medium-to-strong processing strengths. In particular, we considered JPEG compression, WebP compression, median filtering (MeF,



(a) JPEG

(b) WebP

Fig. 5: Accuracy of Baseline (Base.) and RS detector (RS) on compressed images.

kernel size 7 × 7), mean filtering (MF, kernel size 7 × 7), and the print and scan operation (P&S), which affects colors and geometry, simulated using the network in [17]. The MSE introduced by the various processings is in the range $[10^{-3}, 10^{-1}]$. For these experiments we focus on the StyleGAN2 detection task, in the setting with input size 224 × 224, $\sigma = 1$. Fig. 2 shows an example of an image processed with the various post-processing operations.

The accuracy of the RS detectors on post-processed images is reported in Table II and Fig. 5, where it is compared to that achieved by the baseline detectors using the same architecture (that is, trained in the same way of the base models but without the addition of the noise). We see that, especially in the EfficientNet case, the RS detector gets strong robustness against all the types of processing. In particular, a strong robustness against compression, both JPEG and WebP, is achieved by the model, which remains effective also when the images are strongly compressed (QF = 10). We also notice that the limited robustness performance of the CLIP-based RS detector are a consequence of the pretty low detection accuracy of this detector (see discussion in Section IV-B). The baseline CLIP model already exhibits a certain degree of robustness which is a consequence of the inherent robustness of CLIP features. It is worth pointing out that, in most cases, the distortion introduced by the post-processing is larger/much larger than the distortion that the RS detector can *provably* withstand, i.e., the certified MSE. This means that the robustness of the RS detectors remains good also when the distortion significantly exceeds the certified value, although, in principle, with such an amount of distortion, the model can be fooled with a different processing or attack.

V. Conclusions

In this paper, we propose to exploit the RS approach to build synthetic image detectors with certified robustness. Our experiments reveal that the input pre-processing, the input size and, above all, the network architecture adopted to implement the RS detector play a significant role. In particular, a high degree of certification can be achieved using an EfficientNetB4 base detector, with a certified MSE on the order of $10^{-4}/10^{-5}$ for a target detection accuracy of 0.8 in the case of Style-GAN2 detection. Future works will focus on further investigating the impact of the input size and resolution, and trying to improve the performance in the case of CLIP, e.g., by exploiting unsupervised learning to modify the embeddings. Evaluating the practical robustness of the RS detectors by attacking them with adversarial attacks is also an interesting direction of research.

References

[1] N. Carlini and H. Farid, "Evading deepfake-image detectors with white-and black-box attacks," in *Proc.of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 658–659.

[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Int. Conference on Learning Representations*, 2018.

[3] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Int. conference on machine learning*. PMLR, 2018, pp. 274–283.

[4] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Int. conference on machine learning*. PMLR, 2019, pp. 1310–1320.

[5] 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)*. IEEE, 2019, pp. 656–672.

[6] M. Barni, M. Stamm, and B. Tondi, "Adversarial multimedia forensics: Overview and challenges ahead," in *2018 26th European signal processing conference (EUSIPCO)*. IEEE, 2018, pp. 962–966.

[7] U. Ojha, Y. Li, and Y. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proc.of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24480–24489.

[8] D. Cozzolino, G. Poggi, R. Corvi, Matthias Nießner, and Luisa Verdoliva, "Raising the bar of ai-generated image detection with clip," in *Proc.of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4356–4366.

[9] D. Tariang, R. Corvi, D. Cozzolino, G. Poggi, K. Nagano, and L. Verdoliva, "Synthetic image verification in the era of generative artificial intelligence: What works and what isn't there yet," *IEEE Security & Privacy*, 2024.

[10] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proc.of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc.of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.

[12] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc.of the IEEE Int. conference on computer vision*, 2015, pp. 3730–3738.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc.of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Int. conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[15] L. Abady, J. Wang, B. Tondi, and M. Barni, "A siamese-based verification system for open-set architecture attribution of synthetic images," *Pattern Recognition Letters*, vol. 180, pp. 75–81, 2024.

[16] O. Grinchuk, A. Parkin, and E. Glazistova, "3d mask presentation attack detection via high resolution face parts," in *Proc.of the IEEE/CVF Int. Conference on Computer Vision*, 2021, pp. 846–853.

[17] N. Purnekar, L. Abady, B. Tondi, and M. Barni, "Improving the robustness of synthetic images detection by means of print and scan augmentation," in *Proc.of the 2024 ACM Workshop on Information Hiding and Multimedia Security*, 2024, pp. 65–73.