

Self-supervised GAN Detector

Yonghyun Jeong¹, Doyeon Kim¹, Pyounggeon Kim¹, Youngmin Ro¹, Jongwon Choi²
¹Samsung SDS, ²Chung-Ang University

yhyun.jeong, dy31.kim, siru.kim, youngmin.ro@samsung.com, choijw@cau.ac.kr

Abstract

Although the recent advancement in generative models brings diverse advantages to society, it can also be abused with malicious purposes, such as fraud, defamation, and fake news. To prevent such cases, vigorous research is conducted to distinguish the generated images from the real images, but challenges still remain to distinguish the unseen generated images outside of the training settings. Such limitations occur due to data dependency arising from the model's overfitting issue to the training data generated by specific GANs. To overcome this issue, we adopt a self-supervised scheme to propose a novel framework. Our proposed method is composed of the artificial fingerprint generator reconstructing the high-quality artificial fingerprints of GAN images for detailed analysis, and the GAN detector distinguishing GAN images by learning the reconstructed artificial fingerprints. To improve the generalization of the artificial fingerprint generator, we build multiple autoencoders with different numbers of upconvolution layers. With numerous ablation studies, the robust generalization of our method is validated by outperforming the generalization of the previous state-of-the-art algorithms, even without utilizing the GAN images of the training dataset.

1. Introduction

Based on the recent enhancement of the generative models, such as Generative Adversarial Networks (GAN) [1], it has become easy to obtain high-quality fake images generated by deep-learning models [2, 3]. Many recent generative models can even transform the target images to include the specific properties of the users' choices [4–8]. However, as technology progresses, the danger of intentionally misusing the fake images for purposes like fraud, defamation, and fake news grows [9–11]. To prevent such cases, it is important to distinguish between the real images and the generated images by GANs [12].

Previous image forgery detection techniques tended to focus on unnatural features in human faces [13, 14], but many recent models have evolved to uncover distinctive fea-

tures produced during the image generation process [12]. For example, the fingerprints generated during the upsampling estimation of the generator can be successfully detected in the frequency domain [15–18]. Unfortunately, the frequency-level fingerprints differ between generative models and object categories, causing GAN detectors to be data-dependent. As a result, when the GAN detectors are tested with the generated images of the unseen GAN models or object categories, they experience a performance reduction [19]. Although [20, 21] tried to overcome these limitations by using transfer learning with a little amount of data, it is unreasonable to expect to be able to get training data for the unknown GAN models employed by the abusers [22]. For the generalization of GAN detectors, it is far more efficient to use unsupervised or self-supervised training.

Therefore, we suggest a novel self-supervised framework composed of the artificial fingerprint generator and the GAN detector. The artificial fingerprint generator is trained only with the real images to reconstruct the diverse artificial fingerprints of the generated images by GANs. Based on the well-known analysis that the upconvolution layers produce the artifacts [16, 23, 24], we build our artificial fingerprint generator to be composed of multiple autoencoders effectively reproducing the diverse artificial fingerprints of GAN models. To further consider the GAN models where the upconvolution layers are removed [6, 25], we also utilize the additional autoencoder designed to reconstruct the input image without any upsampling operation. Then, the GAN detector is trained to distinguish the real images and the reconstructed images from the artificial fingerprint generator, so we can ignore the fake images of the training datasets in the overall training processes. Unlike unsupervised GAN networks [1], which generate images of varying quality based on learning the data distributions of real images, autoencoders can reconstruct high-quality images that are nearly identical to the original real images while effectively avoiding data dependency due to a specific GAN dataset. Our method is tested with various real-world scenarios to validate the state-of-the-art generalization ability of our model to detect the unseen GAN models and object categories. In

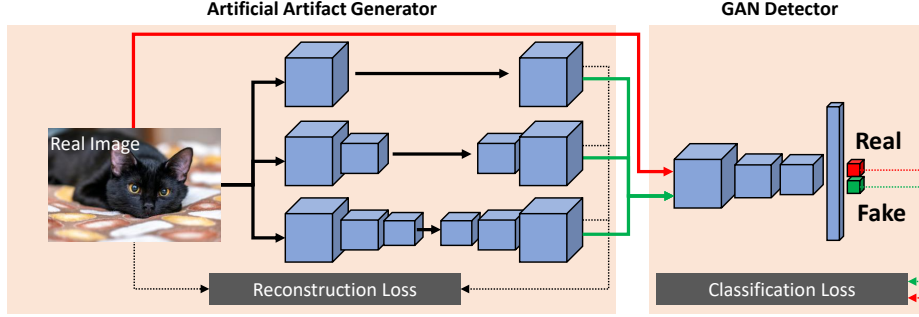


Figure 1. **Overall Framework.** The proposed framework consists of two modules including the artificial fingerprint generator and the GAN detector. The artificial fingerprint generator contains multiple autoencoders trained by the reconstruction loss, while the GAN detector is a conventional classification network to distinguish the real images from the reconstructed images from the artificial fingerprint generator.

addition, our additional experiments reveal that when extra target fake images are included in the GAN detector’s training, our approach shows reasonable improvements compared to the previous studies.

We can summarize our contributions as follows:

- Unlike the previous GAN detectors dependent on the generated images by GAN models, our model utilizes the self-supervised training method to obtain generalized detection ability and avoid data dependency.
- We build the artificial fingerprint generator that replicates the generation of artifacts from the various GAN models independent of their upconvolution layer composition, allowing the GAN detector to be trained robustly across different GAN models and categories.
- Our method shows robust performance compared to the previous GAN detectors even without training the fake images generated by GAN models.

2. Related Work

Image-based detection, frequency-based detection, and transfer learning for detection are the three categories of earlier literature on the detection of created pictures.

2.1. Image-based Detection

To begin, pixel-level characteristics of GAN-based generated images can be employed as identifiers to differentiate between genuine and generated images. The identifiers are referred to as ‘artifacts,’ and they are formed as a result of the generator’s operations in GANs [16, 18, 19]. According to some prior research, discrepancies in blocking artifacts formed during JPEG compression [26, 27] or demosaicing artifacts created by a color filter array [28, 29] should be investigated. Blob-shaped abnormalities, which can be seen in the produced photos of [2], are one of the most well-known pixel-level artifacts. The blob-shaped artifacts exist

in apparent shapes in the generator’s intermediate feature maps, even though they become extremely subtle in the final image [3]. Other image-based detection methods include an adaptable autoencoder-based neural network architecture for new target domains [20] and cross-model manipulation detection using post-processing techniques like JPEG and blur [30].

2.2. Frequency-based Detection

For improved performance, a number of GAN detectors focus on the distinctive patterns in the frequency spectra. [31] proposes combining the variance of the prediction residue with artifacts in the spatial, frequency domain, whereas [32] proposes using the Fast Fourier Transform [33] and singular value decomposition to distinguish image manipulations like JPEG compression, Gaussian noise, and blur. In addition, [34] uses artificial fingerprints to detect forged regions using frequency-based, GAN-specific detection, whereas [35] suggests a manipulation localization utilizing frequency domain correlation and geographical maps. Recently, [15] considers utilizing the Discrete Cosine Transform [36] to analyze GAN-related artifacts in frequency space, while [37] provides a detection approach based on the artifacts caused by the up-sampler of GANs. In addition, [16, 17] suggests using Azimuthal integration to leverage spectral aberrations for detecting in-authentic photos. However, the frequency-level fingerprints are difficult to generalize since the design of the generators can vary their appearance.

2.3. Transfer Learning for GAN Detector

It is impossible to update the model’s training in a supervised manner on a daily basis since new manipulation methods arise on a daily basis [38]. Transfer learning may be used to improve the generalization of GAN detectors to prevent this problem, according to [20, 21, 38]. Transfer learning is the process of using a model that has already been trained for one job to do another task with less

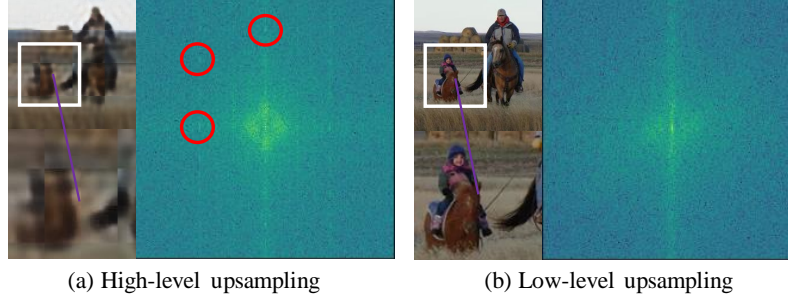


Figure 2. **The reconstructed images of the autoencoders and their 2D spectra.** (a) The high-level upsampling image of the white box and its 2D spectrum with the artificial fingerprints in the red circles, and (b) the low-level upsampling image of the white box and its 2D spectrum. The frequency-level fingerprints appear frequently in high-level upsampling, while the box size of pixel-level checkerboard artifacts is enlarged.

data. Transfer learning can enable the GAN detector to successfully adapt to new domains utilizing a few samples of generated images during training, as proposed by [20] in Forensic-Transfer. To improve detection performance, [21] recently suggested a transferable framework for GAN detectors based on self-training, which successfully recognizes generated images with a little amount of data.

3. Self-supervised GAN Detector

As illustrated in Fig. 1, our framework comprises of two modules: the artificial fingerprint generator and the GAN detector. The artificial fingerprint generator is trained to reconstruct the input images in the unsupervised scheme, while the GAN detector is trained to distinguish the origin real images from the images reconstructed by the artificial fingerprint generator. The artificial fingerprint generator is trained first, and the GAN detector is trained by using the images generated by the trained generator.

3.1. Artificial Fingerprint Generator

As shown in Fig. 2, since the appearance of the frequency-level fingerprints can vary by the number of upsampling operations in the generator network of GAN models, our proposed framework should be able to distinguish the various appearances for generalized detection. Thus, we design the generator using multiple autoencoders with a diverse number of upsampling operations, which is named as the ‘*artificial fingerprint generator*.’ For the deconvolution-based networks, using the autoencoders with the entire scales can enlarge the size of the framework and also slow down the training process; thus, we design an effective framework by mixing up the two artificial fingerprints generated by the large and small number of convolution layers, respectively. Also, to manage the interpolation-based GAN models [6, 25], we employ an additional autoencoder that reconstructs images without the upsampling operations. Thus, the artificial fingerprint generator consists of three autoencoders, which include the autoencoders containing the

high and low levels of deconvolution layers respectively, and an autoencoder without any upsampling operation.

The first autoencoder is the high-level upsampling autoencoder (G_{high}), which applies the downsampling process of $1/64$ to the input images and then applies the upsampling process to reconstruct the original inputs. The second autoencoder is the low-level upsampling autoencoder (G_{low}), which applies the downsampling process of $1/2$ and then applies the upsampling process. The upsampling blocks of the first and second autoencoders are built by a deconvolution layer and the activation function of ReLU [39]. Since the layer of the autoencoder for downsampling is proceeded with the stride 2, t times of downsampling for $2^t = d$ must be proceeded to reach the size of $1/d$. Also, since upsampling is proceeded with stride 2, t times of upsampling must also be proceeded. Thus, G_{high} contains an encoder of 6 convolution layers and a decoder of 6 transposed convolution layers, while only one convolution layer and one transposed convolution layer are used in G_{low} .

The third autoencoder is the non-upsampling autoencoder (G_{non}), which is to detect the images without applying the deconvolution process. To avoid applying the deconvolution process, only the convolution layer is used, and the strides are set to 1 to avoid upsampling operations. G_{non} is composed of two convolution layers respectively for the encoder and the decoder.

The training set for the three autoencoders consists of the real images only. The training loss is a reconstruction loss using the mean squared error as:

$$\mathcal{L}(X, G) = \mathbb{E}_{x \sim X} [\|x - G(x)\|_2^2], \quad (1)$$

where X is the training set containing the samples x and G is one of the autoencoders including G_{high} , G_{low} , and G_{non} . The three autoencoders are individually trained using the same datasets, but they generate different types of artificial fingerprints due to various levels of upsampling operations.

		Acc.	A.P.
Supervised	Wang [30]	50.4	63.8
	Frank [15]	78.9	77.9
	Durall [16]	85.1	79.5
Unsupervised	Ours	92.0	97.7

Table 1. Cross-category performance.

3.2. GAN Detector

After finishing the training of the artificial fingerprint generator, we train the GAN detector to distinguish the generated images from the real images. The frequency-level analysis is effective in observing the artificial fingerprints. Thus, we also utilize the 2D spectrum by Fast Fourier Transform (FFT) [33]. The generated image $\hat{x} \in \hat{X} = \{G_{high}(x), G_{low}(x), G_{non}(x) | \forall x \in X\}$ of the artificial fingerprint generator is transformed into a 2D spectrum by FFT.

For the fair comparison with the recent models [15, 16, 30], we design our GAN detector based on ResNet-50 [40]. To train the GAN detector, we reconstruct all real images of the training dataset, which is considered as the generated images. Since the three autoencoders reconstruct three images from a real image, the training of GAN detector can suffer from the unbalancing problem. Thus, when we compose the mini-batch, the sampling probability is adjusted to extract the real images three times more frequently than the reconstructed images.

In addition to the raw images, we utilize the mixup algorithm [41] to consider the mixed artificial fingerprints by integrating all three types of the artificial fingerprints with the real images. When we define two samples randomly selected from a mini-batch ($\mathbf{S} = \{X, \hat{X}\}$) by \mathbf{s}_i and \mathbf{s}_j and their corresponding one-hot labels are \mathbf{y}_i and \mathbf{y}_j , respectively, the mixed sample $\tilde{\mathbf{s}}_{(i,j)}$ and its corresponding label $\tilde{\mathbf{y}}_{(i,j)}$ are obtained by:

$$\tilde{\mathbf{s}}_{(i,j)} = \lambda \mathbf{s}_i + (1 - \lambda) \mathbf{s}_j, \quad \tilde{\mathbf{y}}_{(i,j)} = \lambda \mathbf{y}_i + (1 - \lambda) \mathbf{y}_j, \quad (2)$$

where $\lambda \in [0, 1]$ is a mixing scale randomly chosen by a Beta distribution.

Then, we replace every sample of the mini-batch with the mixed samples by iterating the mixup algorithm, which is defined by $\tilde{\mathbf{S}}$. The labels corresponding to the samples of $\tilde{\mathbf{S}}$ is defined as $\tilde{\mathbf{Y}}$. Then, the GAN detector (C) is trained by a conventional softmax cross-entropy loss as:

$$\mathcal{L}_C(\tilde{\mathbf{S}}) = \mathbb{E}_{(\tilde{\mathbf{s}}, \tilde{\mathbf{y}}) \sim (\tilde{\mathbf{S}}, \tilde{\mathbf{Y}})} [\tilde{\mathbf{y}}^T \log(C(\tilde{\mathbf{s}}))]. \quad (3)$$

4. Experimental Results

4.1. Dataset

Through experiments, we compare the performance of each network by employing the same data. Since the training settings have a strong impact on the analysis of the GAN detector, we utilize the real horse images of LSUN [43], which is the dataset used for the training of ProGAN [44]. In contrast, the comparing models use the generated images of horse, car, cat, and airplane categories of ProGAN [44], as well as the real images of LSUN [43]. We utilize the 20 different object categories of ProGAN [44] to observe the changes in the performance of the GAN detectors based on the change in categories of the same model.

Also, to evaluate the level of data dependency due to a specific GAN dataset, we utilize the most well-known unconditional GAN models with high resolutions, such as ProGAN [44], StyleGAN [2], and StyleGAN2 [3]. Then, to evaluate the performance of the GAN detectors on generating new images from the latent space, we employ the conditional GAN model named BigGAN [25], which is known to realistically reconstruct the data of the most categories, and the image-to-image translation models, such as CycleGAN [4], StarGAN [5], and GauGAN [45]. Lastly, we conduct experiments to test the models whether they can detect not only the entire synthesis of face images but also partially generated images using FaceSwap [46]. We utilize various GANs with human faces and various objects, and the real images used to train the GANs, including CelebA-HQ [47], CelebA [48], COCO [49], LSUN [43], and ImageNet [50], which are all publicly available through [30] and [21].

4.2. Evaluation Metrics

For performance comparison, we employ the accuracy and average precision [51], which are the metrics commonly used in this field of study. To test the transfer-learning performance, we use the Area Under the Receiver Operating Characteristic (AUROC) as used in [20, 21]. To compare the generalization performance, we follow the suggestion of Wang *et al.* [30] to use JPEG compression, which is known as the most effective method to test the generalization performance. Also, for the frequency-level analysis, we compare with Frank *et al.* [15] and Durall *et al.* [17], both of which study the frequency-based detection. Lastly, for zero-shot and domain transfer performance, we compare with Jeon *et al.* [21] and Cozzolino *et al.* [20].

4.3. Implementation Details

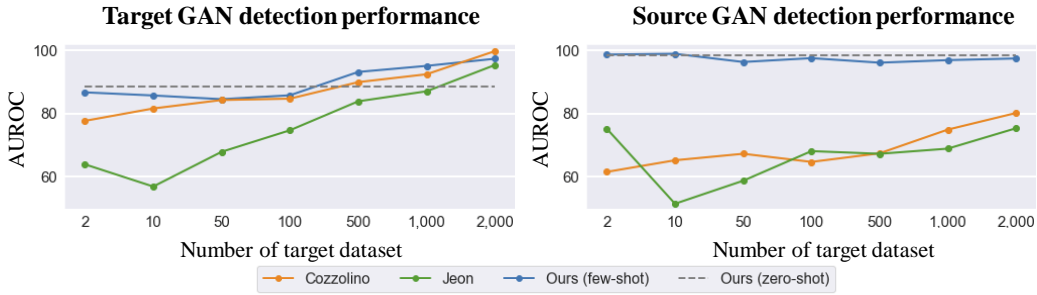
We resize the input images into 256×256 , and the reconstructed images with the same size. For training, we use a single NVIDIA RTX 8000 with a batch size of 16 and 20 epochs for the GAN detector, 80 epochs for the artificial fingerprint generator. Both of the artificial fingerprint

Table 2. **Zero-shot cross-model performance.**

Model	StyleGAN		StyleGAN2		BigGAN		CycleGAN		StarGAN		GauGAN		Deepfake		Mean	
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
Wang [30]	63.8	91.4	76.4	97.5	52.9	73.3	72.7	88.6	63.8	90.8	63.9	92.2	51.7	62.3	63.6	85.2
Frank [15]	72.2	82.1	64.2	80.1	68.9	82.4	53.7	66.2	89.1	99.2	65.3	90.3	51.1	49.6	66.4	78.6
Durall [16]	63.9	58.4	69.0	62.7	58.5	54.7	69.6	63.1	99.0	98.1	57.0	53.8	50.4	50.1	66.8	63.0
Ours	71.5	79.1	70.0	79.3	77.3	90.0	57.5	86.3	99.8	100.	70.9	96.9	69.2	74.0	73.7	86.5

Table 3. **Zero-shot cross-model performance.**

Model	number of training class (real/fake)	StyleGAN		StyleGAN2		BigGAN		CycleGAN		StarGAN		GauGAN		Deepfake		Mean	
		Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
Wang [42]	(1,1)	50.4	79.3	68.2	94.7	50.2	61.3	50	52.9	50	58.2	50.3	67.6	50.1	51.5	52.7	66.5
Frank [15]	(1,1)	68.5	80.7	60.8	77.3	72.1	63	57.6	56.6	80.1	76.3	74	95.5	54.4	59.4	66.8	72.7
Durall [16]	(1,1)	64.1	58.6	69.3	62.9	55.4	52.9	69.6	62.8	95.4	91.5	57.5	54	53.6	52	66.4	62.1
Wang [42]	(2,2)	52.8	82.8	75.7	96.6	51.6	70.5	58.6	81.5	51.2	74.3	53.6	86.6	50.6	51.5	56.3	77.7
Frank [15]	(2,2)	70.8	83.8	61.2	75.6	74.9	76.2	74.8	76.8	91.7	97.5	89.2	98.4	52.8	53	73.6	80.2
Durall [16]	(2,2)	63.5	58.1	68.7	62.4	56.4	53.5	63.5	58.2	89.8	83.1	56.5	53.5	53.3	51.7	64.5	60.1
Wang [42]	(4,4)	63.8	91.4	76.4	97.5	52.9	73.3	72.7	88.6	63.8	90.8	63.9	92.2	51.7	62.3	63.6	85.2
Frank [15]	(4,4)	72.2	82.1	64.2	80.1	68.9	82.4	53.7	66.2	89.1	99.2	65.3	90.3	51.1	49.6	66.4	78.6
Durall [16]	(4,4)	63.9	58.4	69	62.7	58.5	54.7	69.6	63.1	99	98.1	57	53.8	50.4	50.1	66.8	63.0
Wang [42]	(20,20)	71.4	96.3	67.5	93.4	60.9	83.3	83.8	94.3	84.6	93.6	79.3	98.1	51.1	66.3	71.2	89.3
Frank [15]	(20,20)	81.8	91.7	71.4	93.0	76.0	87.8	62.8	77.3	96.9	99.4	73.9	93.1	48.6	48.3	73.1	84.4
Durall [16]	(20,20)	64.7	59.0	69.2	62.9	59.4	55.3	66.9	60.9	98.5	97.1	57.2	53.9	52.2	51.3	66.9	62.9
Our	(1,0)	71.5	79.1	70.0	79.3	77.3	90.0	87.5	86.3	99.8	100.	70.9	96.9	69.2	74.0	78.0	86.5

Figure 3. **Performance comparison on transfer learning.** The performance comparison of GAN detectors based on transfer learning is shown according to a gradual increase of training data for the target and the source, respectively. The target GAN performance is shown on the left, and the source GAN performance is shown on the right.

generator and GAN detector networks are trained by Adam optimizer [42] with the learning rate of 0.0001, which are the conventional hyperparameters of the previous GAN detectors.

4.4. Generalization Performance of GAN Detector

To show the generalization performance of the GAN detectors, we conduct two experiments. First, using the generated images of the same GAN model, we train the GAN detectors with only one object category and test with the entire object categories to evaluate the generalization performance. Table 1 shows the test results using the 20 classes of the GAN detectors trained with the horse category of ProGAN [44]. Our model shows the highest results, even without learning any generated images of ProGAN and other models.

Table 2 shows the results of the second experiment to test the cross-GAN performance of the GAN detectors. We

compare with the previous studies used for the comparison of cross-GAN performance: Wang *et al.* [30], Frank *et al.* [15], and Durall *et al.* [16]. Each of them is trained with the horse, car, cat, and airplane categories of ProGAN [44], and tested with the generated images of seven other generative models. In contrast, our self-supervised GAN detector is trained with the real horse images only. Even with the highly limited setting, our GAN detector achieves the highest accuracy and average precision, even without using any generated images of GAN models.

To show the outstanding performance of our framework, we perform additional experiments with various settings of compared algorithms. As presented in Table 3, we variate the number of categories for training the other generated image detectors. From the results, our framework shows state-of-the-art performance even while using only one training category and no fake images.

Table 4. Few-shot learning performance.

Model	Shot	Test Models															
		StyleGAN		StyleGAN2		BigGAN		CycleGAN		StarGAN		GauGAN		Deepfake		Mean	
		SRC	TRG	SRC	TRG	SRC	TRG	SRC	TRG	SRC	TRG	SRC	TRG	SRC	TRG	SRC	TRG
Cozzolino [20]	1,000	96.8	99.2	71.3	76.1	76.1	94.0	68.5	99.2	62.6	100.	84.0	99.9	65.3	78.2	74.9	92.3
Cozzolino [20]	2,000	98.8	99.7	96.0	99.9	82.6	99.8	65.2	99.9	62.4	100.	76.7	99.9	79.5	98.4	80.1	99.6
Jeon [21]	1,000	67.7	94.0	60.0	89.5	59.7	66.2	71.0	82.5	80.2	99.3	94.6	97.2	49.3	80.2	68.9	86.9
Jeon [21]	2,000	64.3	96.8	69.3	99.4	64.2	85.8	98.3	98.2	85.4	99.8	96.7	98.4	49.5	88.6	75.3	95.2
Ours	0	98.3	76.5	98.3	80.3	98.3	91.9	98.3	89.1	98.3	100.	98.3	97.6	98.3	74.7	98.3	88.6
Ours	1,000	97.9	94.3	97.3	95.0	96.7	98.6	95.0	96.8	98.1	100.	95.5	99.8	97.6	80.6	96.9	95.0
Ours	2,000	98.0	97.3	98.6	98.4	96.6	99.1	95.1	99.4	97.9	100.	97.3	99.8	98.3	86.9	97.4	97.3

Table 5. Ablation study for self-supervised GAN detector.

Model	Test Models																	
	ProGAN		StyleGAN		StyleGAN2		BigGAN		CycleGAN		StarGAN		GauGAN		Deepfake		Mean	
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
Ours	94.1	98.3	71.5	79.1	70.0	79.3	77.3	90.0	57.5	86.3	99.8	100.	70.9	96.9	69.2	74.0	76.3	88.0
w/o G_{high}	92.4	98.2	67.2	77.8	72.6	80.0	71.5	91.1	53.9	88.9	99.7	100.	63.5	84.9	55.3	58.3	72.0	84.9
w/o G_{low}	90.1	97.9	68.1	81.6	61.5	70.2	66.2	77.3	52.0	47.6	98.7	99.9	67.1	83.3	52.9	53.7	69.6	76.4
w/o G_{non}	54.8	88.6	55.0	75.1	59.3	85.3	85.9	92.0	80.0	94.0	100.	100.	85.0	95.7	50.1	56.8	71.3	85.9
w/o Freq.	91.1	98.0	71.3	80.3	67.8	77.8	73.0	77.1	61.4	90.5	99.4	100.	71.3	94.8	65.4	70.6	75.1	86.1
w/o Mixup	92.1	97.6	67.7	76.8	66.2	71.4	70.9	78.8	60.2	87.4	93.7	100.	69.4	93.4	56.0	59.6	72.0	83.1

4.5. Transfer GAN Domain Performance

In this section, we compare the zero-shot performance of our model and other models employing transfer learning, which is also known as few-shot learning. The comparing models follow the standards of [20, 21] to utilize the 1,000 to 2,000 generated images of a specific GAN model for transfer learning, and ProGAN [44] as the source model. To apply transfer learning to our model, after finishing the baseline training of our GAN detector, we fine-tune the model by 1 epoch with the mixed datasets composed of the additional target dataset and the reconstructed data. The proportion of the target GAN domain in the entire batch is set to 15%.

Fig. 3 shows the performance comparison of the models trained with 2 to 2,000 real and generated images to observe whether the models maintain the performance in the target and source GAN as the number of training images increase. As shown in Table 4, our model not only achieves competitive performance compared to other models employing transfer learning in the target domain but also outperforms others in the source domain by additionally using the reconstructed images of the artificial fingerprint generator. Overall, our model maintains robust performance in the source GAN while achieving increased performance in the target GAN. In the experiment of few-shot learning with the training images under 1,000, our model achieves the most robust performance, and when the training images are expanded to 2,000, our model achieves 97.4% in AUROC.

4.6. Ablation Study

Table 5 shows the individual effect of each component of our proposed method. In the second row, ‘w/o G_{high} ’ indicates the results of downsampling to 1/64 without ap-

plying the artificial fingerprint generator for upsampling. The third and fourth columns also show the results without applying G_{low} , G_{non} . Also, it shows the results of using the pixel-level analysis instead of the frequency-level, and the results without the mix-up algorithm. By excluding the data without the downsampling process, the performances of ProGAN, StyleGAN, and StyleGAN2 dramatically decline, while those of BigGAN, CycleGAN, and GauGAN increase. This indicates the artificial fingerprints generated by the transposed convolution are relatively few in the images of BigGAN, CycleGAN, and GauGAN. Thus, in order to generally detect images generated by various GAN models, it is important to utilize the images reconstructed from the autoencoder without the upsampling process. Also, the mixup algorithm enables the artificial fingerprint generator to generate various sizes of artificial fingerprints, which boosts up the training effect. Interestingly, the achievement from the frequency-level transformation seems minor compared to the other components, allowing us to claim that the artificial fingerprints can be detected effectively even with the pixel-level images.

5. Conclusion

We propose a self-supervised GAN detector with robust generalization ability to detect the generated images even upon the unseen setting during training. Through a comprehensive analysis of the frequency-level fingerprints found in the previous studies, we build the integrated framework composed of the artificial fingerprint generator and GAN detector. By reconstructing the artificial fingerprints in the frequency domain and utilizing them as the training data, our proposed method achieves state-of-the-art performance even without using any generated images by GANs. Especially, our method shows impressive performance when

tested with unseen categories and GAN models, which validates the robust generalization of our GAN detector. Based on the various frequency-level fingerprints of the generated images, we plan to analyze the missing characteristics of the fingerprints to further extend our framework.

References

- [1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 1
- [2] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1, 2, 4
- [3] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. *CoRR*, abs/1912.04958, 2019. 1, 2, 4
- [4] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017. 1, 4
- [5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 4
- [6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 3
- [7] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14104–14113, 2020. 1
- [8] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European conference on computer vision*, pages 592–608. Springer, 2020. 1
- [9] Sangyup Lee, Shahroz Tariq, Youjin Shin, and Simon S Woo. Detecting handcrafted facial image manipulations and gan-generated facial images using shallow-fakefacenet. *Applied Soft Computing*, 105:107256, 2021. 1
- [10] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deep-fake detection dataset. *arXiv preprint arXiv:2103.10094*, 2021. 1
- [11] Thanh Thi Nguyen, Cuong M Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. Deep learning for deepfakes creation and detection. *arXiv preprint arXiv:1909.11573*, 2019. 1
- [12] Ke Sun, Hong Liu, Qixiang Ye, Jianzhuang Liu, Yue Gao, Ling Shao, and Rongrong Ji. Domain general face forgery detection by learning to weight. 2021. 1
- [13] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses, 2018. 1
- [14] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020. 1
- [15] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning*, pages 3247–3258. PMLR, 2020. 1, 2, 4, 5
- [16] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your Up-Convolution: CNN Based Generative Deep Neural Networks are Failing to Reproduce Spectral Distributions. In *IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, United States, 2020. 1, 2, 4, 5
- [17] Ricard Durall, Margret Keuper, Franz-Josef Pfrendt, and Janis Keuper. Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686*, 2019. 1, 2, 4
- [18] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. *arXiv preprint arXiv:2105.02577*, 2021. 1, 2
- [19] Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are gan generated images easy to detect? a critical analysis of the state-of-the-art. *arXiv preprint arXiv:2104.02617*, 2021. 1, 2
- [20] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv*, 2018. 1, 2, 3, 4, 6
- [21] Hyeonseong Jeon, Young Oh Bang, Junyaup Kim, and Simon Woo. T-gd: Transferable gan-generated images detection framework. In *International Conference on Machine Learning*, pages 4746–4761. PMLR, 2020. 1, 2, 3, 4, 6
- [22] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European Conference on Computer Vision*, pages 101–117, 2018. 1
- [23] Tarik Dzanic, Karan Shah, and Freddie D. Witherden. Fourier spectrum discrepancies in deep network generated images. In *Advances in Neural Information Processing Systems*, 2020. 1
- [24] Keshigeyan Chandrasegaran, Ngoc-Trung Tran, and Ngai-Man Cheung. A closer look at fourier spectrum discrepancies for cnn-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1
- [25] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 1, 3, 4

- [26] Shuiming Ye, Qibin Sun, and Ee-Chien Chang. Detecting digital image forgeries by measuring inconsistencies of blocking artifact. In *IEEE International Conference on Multimedia and Expo*, pages 12–15. Ieee, 2007. 2
- [27] Dijana Tralic, Juraj Petrovic, and Sonja Grgic. Jpeg image tampering detection using blocking artifacts. In *International Conference on Systems, Signals and Image Processing*, pages 5–8. IEEE, 2012. 2
- [28] Ahmet E. Dirik and Nasir Memon. Image tamper detection based on demosaicing artifacts. In *2009 16th IEEE International Conference on Image Processing*, pages 1497–1500, 2009. 2
- [29] Pasquale Ferrara, Tiziano Bianchi, Alessia De Rosa, and Alessandro Piva. Image forgery localization via fine-grained analysis of cfa artifacts. *IEEE Transactions on Information Forensics and Security*, 7(5):1566–1577, 2012. 2
- [30] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot...for now. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 4, 5
- [31] Matthias Kirchner. Fast and reliable resampling detection by spectral analysis of fixed linear predictor residue. In *ACM workshop on Multimedia and security*, pages 11–20, 2008. 2
- [32] Deng-Yuan Huang, Ching-Ning Huang, Wu-Chih Hu, and Chih-Hung Chou. Robustness of copy-move forgery detection under high jpeg compression artifacts. *Multimedia Tools and Applications*, 76(1):1509–1530, 2017. 2
- [33] James W. Cooley, Peter A.W. Lewis, and Peter D. Welch. The fast fourier transform and its applications. *IEEE Transactions on Education*, 1969. 2, 4
- [34] Francesco Marra, Diego Gagnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *IEEE Conference on Multimedia Information Processing and Retrieval*, pages 506–511. IEEE, 2019. 2
- [35] Jawadul H Bappy, Cody Simons, Lakshmanan Nataraj, BS Manjunath, and Amit K Roy-Chowdhury. Hybrid lstm and encoder-decoder architecture for detection of image forgeries. *IEEE Transactions on Image Processing*, 28(7):3286–3300, 2019. 2
- [36] Nasir Ahmed, T. Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974. 2
- [37] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *IEEE International Workshop on Information Forensics and Security*, pages 1–6, 2019. 2
- [38] Shivangi Aneja and Matthias Nießner. Generalized zero and few-shot transfer for facial forgery detection. *arXiv preprint arXiv:2006.11863*, 2020. 2
- [39] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010. 3
- [40] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 4
- [41] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2018. 4
- [42] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014. 5
- [43] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 4
- [44] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 4, 5, 6
- [45] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 4
- [46] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *IEEE International Conference on Computer Vision*, pages 1–11, 2019. 4
- [47] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 4
- [48] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, December 2015. 4
- [49] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4
- [50] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 4
- [51] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2010. 4