Towards Robust Gan-Generated Image Detection: A Multi-View Completion Representation

Chi Liu[1] , **Tianqing Zhu**[1] , **Sheng Shen**[2] and **Wanlei Zhou**[3]
[1]School of Computer Science, University of Technology Sydney, Australia
[2]School of Electrical and Information Engineering, The University of Sydney, Australia
[3]City University of Macau, Macao SAR, China
tianqing.zhu@uts.edu.au

Abstract

GAN-generated image detection now becomes the first line of defense against the malicious uses of machine-synthesized image manipulations such as deepfakes. Although some existing detectors work well in detecting clean, known GAN samples, their success is largely attributable to overfitting unstable features such as frequency artifacts, which will cause failures when facing unknown GANs or perturbation attacks. To overcome the issue, we propose a robust detection framework based on a novel multi-view image completion representation. The framework first learns various view-to-image tasks to model the diverse distributions of genuine images. Frequency-irrelevant features can be represented from the distributional discrepancies characterized by the completion models, which are stable, generalized, and robust for detecting unknown fake patterns. Then, a multi-view classification is devised with elaborated intra- and inter-view learning strategies to enhance view-specific feature representation and cross-view feature aggregation, respectively. We evaluated the generalization ability of our framework across six popular GANs at different resolutions and its robustness against a broad range of perturbation attacks. The results confirm our method's improved effectiveness, generalization, and robustness over various baselines.

1 Introduction

AI-powered image manipulation techniques, such as deepfakes, are constantly evolving thanks to the continuous advances in deep generative models, particularly generative adversarial networks (GANs) [Goodfellow *et al.*, 2014a]. The quality and fidelity of the generated images have reached a photorealistic level that is indistinguishable from real images to human eyes. Alongside the technical advance, society is raising significant concerns regarding the abuse of these techniques to create and spread misleading information, which will cause a trust crisis where "seeing is no longer believing." To tackle the issues, the research community has been dedicated to developing powerful forensics tools against malicious image manipulations. One crucial and promising direction is detecting GAN-generated fake images, considering the ubiquitous adoptions of GANs in image manipulation tasks.

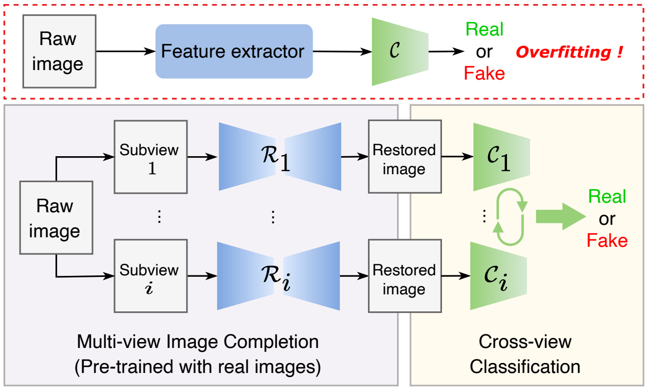

Figure 1: Instead of learning GAN-specific features directly from fake images which may lead to overfitting, our framework incorporates multi-view completion and classification to model diverse distributional discrepancies between real and fake images, which can generalize to unknown fake patterns. $\mathcal{R}$: Restorer; $\mathcal{C}$: Classifier.

Most detection methods typically train CNN classifiers to learn specific features to distinguish GAN-generated images from real ones [Hu *et al.*, 2021; Liu *et al.*, 2020; Marra *et al.*, 2019a; Dzanic *et al.*, 2020; Frank *et al.*, 2020; Durall *et al.*, 2020], which work satisfactorily on clean test samples from the same GANs used in training. However, their performances will decrease dramatically when facing samples generated by unknown GANs or perturbed by noises, leading to limited practical reliability. One primary reason is that a deep CNN classifier may easily overfit unstable GAN-specific features of the training samples, particularly the low-level frequency-domain artifacts [Wang *et al.*, 2020; He *et al.*, 2021; Jeong *et al.*, 2022a; Jeong *et al.*, 2022b]. Previous studies have proved that conspicuous artifacts exist in the spectra of GAN-generated images [Wang *et al.*, 2020; Frank *et al.*, 2020]: Despite being easily identified by classifiers, these artifact patterns are inconsistent, varying significantly among different GAN models or perturbations. As a result, the classifier overfitting a specific frequency pattern will suffer from weak generalization ability and robustness in detecting other frequency patterns.

Based on the understanding of the overfitting issue, we are motivated to design an improved detection model with two requirements: (1) reduce the dependency on unstable low-level frequency features; and (2) learn a robust feature representation from other types of information, such as regional consistency, and color or textural details of images. Instead of directly learning detectable features from fake images, which potentially leads to the frequency overfitting problem, we propose a novel detection framework that incorporates a multi-view image completion learning and a cross-view classification learning processes, as sketched in Fig. 1. The framework can learn a strong and stable feature representation from diverse frequency-independent, view-specific information, resulting in outperforming generalization and robustness when facing unknown GANs or perturbations.

In the multi-view completion process, multiple view-to-image completion models are learned *with real images only*, and then used to characterize diverse distributional discrepancies between real and fake images. In contrast to overfitting specific GAN patterns, the compact distributions of the image-missing characteristics modeled from real images are more likely to distinguish unknown, out-of-distribution fake images from real ones [Ruff *et al.*, 2020]. In addition, the view-to-image completions can align the frequency patterns of different types of fake samples with that of real images, which helps reduce the classifier's frequency bias. Then, in the cross-view classification, the real and fake samples synthesized from each incomplete view are fed into an independent classifier. The multi-scale feature concatenation and low-pass residual-guided attention modules are devised to strengthen the intra-view feature representation. The independent classifiers are finally combined using an adaptive loss fusion strategy to enhance the learning from cross-view information. Our contributions are highlighted as follows:

- We propose a novel GAN-generated image detection framework using multi-view completion classification learning to build a robust feature representation for detecting unknown GANs and perturbations.
- We devise several novel modules and learning strategies that effectively benefit the framework's ability to capture and incorporate diverse view-specific features.
- We perform extensive evaluations which validate the significantly improved generalization and robustness of our framework in a wide range of settings varying in image resolutions, GAN types, and perturbation methods.

2 Generated Image Detection: A Review

Image-domain detection. Image-domain detection extracts detectable traces from the pixel inputs. Earlier works tended to train a CNN to learn deep features in a data-driven manner [Marra *et al.*, 2018; Tariq *et al.*, 2018], while more recent works prefer to craft specific features for higher detection accuracy, such as the co-occurrence matrices [Nataraj *et al.*, 2019; Barni *et al.*, 2020], saturation [McCloskey and Albright, 2019], specular highlights [Hu *et al.*, 2021] and texture cues [Liu *et al.*, 2020]. [Marra *et al.*, 2019a] and [Yu *et al.*, 2019] pointed out that a GAN will leave a unique fingerprint containing the source model information in the generated images. Some other works improve on the network design, where novel learning strategies or modules, such as incremental learning [Marra *et al.*, 2019b], self-attention mechanism [Jeon *et al.*, 2020; Mi *et al.*, 2020] and vision transformer [Wang *et al.*, 2022], are adopted.

Frequency-domain detection. Frequency-domain detection relies on identifying the frequency discrepancy between GAN-generated images and real images [Dzanic *et al.*, 2020; Frank *et al.*, 2020; Durall *et al.*, 2020]. Frequency discrepancy can be easily captured in various spectral representations by a classifier. For example, [Frank *et al.*, 2020] found that even a shallow CNN can achieve a high detection accuracy using the 2D Discrete Cosine Transform (DCT) coefficients as input data. [Qian *et al.*, 2020] proposed a dual-branch network that extracts the global and local DCT features. [Dzanic *et al.*, 2020] and [Durall *et al.*, 2020] proposed to transform the 2D Fast Fourier Transform (FFT) magnitude into 1D power profile as detectable features. [Liu *et al.*, 2021] found that more distinguishable features can be extracted in the phase spectrum than in the amplitude spectrum. Unfortunately, some recent studies also pointed out that the frequency features are unstable and easy to be concealed [Dzanic *et al.*, 2020; Durall *et al.*, 2020; Liu *et al.*, 2022; Huang *et al.*, 2020; Jung and Keuper, 2021; Dong *et al.*, 2022]. Thus, detectors heavily relying on frequency features are vulnerable and weakly generalized.

Generalized and robust detection. Generalized and robust detection of GAN-generated images is now in high demand. Most existing works involve a preprocessing operation to strengthen the representation of generalized and robust features. [Zhang *et al.*, 2019] pointed out that a detector can generalize between two GANs with similar spectral artifacts in their generated images, which in turn confirms that a detector is likely to overfit specific frequency patterns. [Wang *et al.*, 2020] explored the effects of different data augmentation strategies such as compression and blurring in improving a detector's generalization ability. [Jeong *et al.*, 2022a] proposed to preprocess the fake images with a bilateral high-Pass filter, which amplifies the effect of the common frequency-level artifacts shared by different GANs. [Jeong *et al.*, 2022b] designed a frequency-level perturbation framework to erode the GAN-specific spectral artifacts in generated images before feeding them to a detector. [He *et al.*, 2021] proposed to re-synthesize training images using a super-resolution model pre-trained with real images to help extract robust features and isolate fake images.

3 The Proposed Framework

We design the **M**ulti-view **C**ompletion **C**lassification **L**earning (MCCL) to build a novel multi-view, frequency-independent feature representation for robust detection of GAN-generated images. As shown in Figure 2, the framework jointly trains a set of restorers and classifiers. The restorers are trained with real images only, and each learns to reconstruct the full image from one particular incomplete view. Then, both real and fake images are processed by each restorer through the same view-to-image pipeline. Since

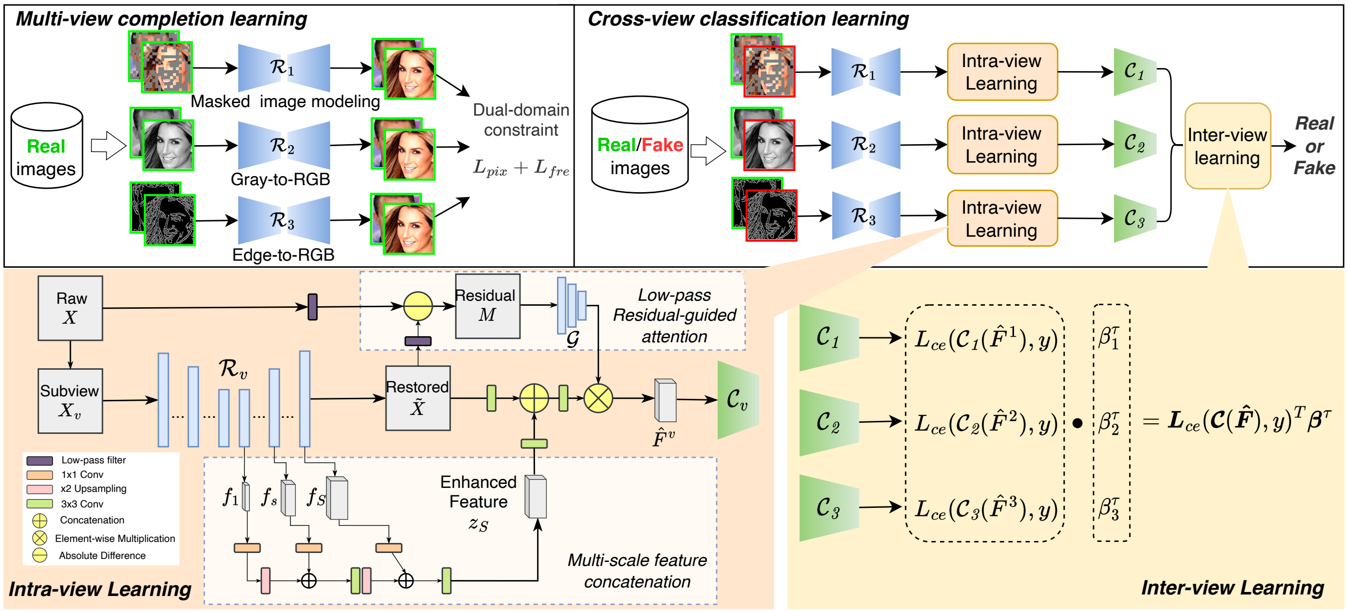

Figure 2: The overview of our framework (white box). Several restorers first learn different distributions of real images via multi-view completion learning. Then for each view, a classifier captures the view-specific distributional discrepancy between real and fake images via intra-view learning. The low-pass residual-guided attention and multi-scale feature concatenation modules are devised to strengthen intra-view learning (orange box). All base classifiers are finally fused to perform inter-view learning for robust detection (yellow box).

the recovery of missing information is governed by real images' characteristics only, the distributional difference between the reconstructed real and fake samples can be reflected in the restored information. Then, for each view, a classifier is trained based on the reconstructed samples to capture the view-specific distributional discrepancy. We combine the multi-scale features encoded by different decoding layers of each restorer with the restored image as the classifier's input. A low-pass residual-guided attention module is employed at the entry of the classifier to highlight the reconstruction difference between real and fake images. A self-adaptive loss fusion module is additionally designed to combine the decisions of multiple classifiers to facilitate inter-view learning.

3.1 Multi-View Image Completion Learning

Several independent encoder-decoder-based restorers $\boldsymbol{\mathcal{R}} = \{\mathcal{R}^v\}_{v=1}^N$ are trained with real images to recover the full image from different incomplete views. It is non-trivial to select the appropriate views for completion, which determines what types of frequency-irrelevant information we want to exploit. Since regional consistency, color, and texture have been proven to be distinguishable features for GAN-generated images [Liu *et al.*, 2020; Hu *et al.*, 2021], we empirically consider three completion tasks: Masked Image Modeling, Gray-to-RGB, and Edge-to-RGB, where the regional, color and textural details are previously missing and restored, respectively. The natural compact distributions of these types of information can be modeled during the completion. The significance of each view is also explored in the experiment.

Masked Image Modeling is an emerging approach for visual representation learning [He *et al.*, 2022], which masks a portion of an image and predicts the masked area, and can be leveraged to model the regional consistency of natural images. The masking strategy is that, given an image $X \in \mathbb{R}^{w\times h\times 3}$, we randomly mask 50% non-overlapping patches with a patch size of $(\frac{w}{16}, \frac{h}{16})$. Gray-to-RGB aims to learn color information from real images. We first transform the RGB image into the gray-scale version and then predict the raw RGB pixel values from the gray-scale input. Edge-to-RGB aims to learn textural information from real images. We first extract the binary edge sketch from the RGB image using the Canny edge detector, and then predict the raw RGB pixel values from the edge input. Figure 3 shows an example of different incomplete views.

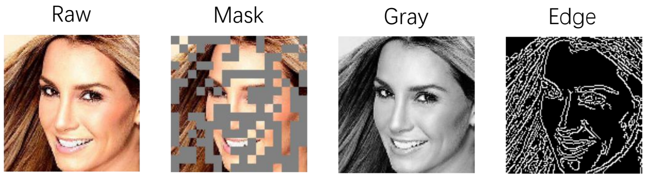

Figure 3: Three incomplete views selected for completion learning.

Dual-domain constraint. Given an image X and an incomplete view X^v, the completion is formulated as $\tilde{X}^v = \mathcal{R}^v(X^v)$. The training of $\mathcal{R}^v$ is supervised with a dual-domain constraint, which incorporates a pixel-level regression loss and a frequency loss:

$$\begin{aligned} L_{pix} &= ||X - \tilde{X}^v||_1 = ||X - \mathcal{R}^v(X^v)||_1, \\ L_{fre} &= ||\mathcal{F}(X) - \mathcal{F}(\tilde{X}^v)||_2^2. \end{aligned} \tag{1}$$

where $\mathcal{F}(\cdot)$ denotes the 2D FFT. The frequency loss computes the element-wise Euclidean distance between the Fourier spectra of original and restored images, which ensures $\mathcal{R}^v$ to capture the natural, correct frequency property of real images, so as to facilitate the frequency alignment between real and fake images. The optimization function can be therefore denoted as

$$\min L_{pix}^v + \lambda L_{fre}^v, \tag{2}$$

where λ is the weight to balance different losses.

3.2 Intra-View Classification Learning

After training $\mathcal{R}^v$ with real images, both real and fake images are processed by $\mathcal{R}^v$ via the same image completion pipeline to enable the subsequent classification learning. To mine more robust and frequency-irrelevant features from each individual view's pathway, we propose the multi-scale feature concatenation and low-pass residual-guided attention modules, as shown in the orange box in Figure 2.

Multi-scale feature concatenation. Since $\mathcal{R}^v$ is an encoder-decoder consisting of multiple layers, during the completion, the missing information of the original image is progressively recovered by the stacked decoding layers of $\mathcal{R}^v$. Thus, meaningful features for distinguishing real and fake images are embedded not only in the final output image, but also in the intermediate feature maps of the decoder. To this end, we build a feature pyramid to concatenate the intermediate features at different scales. For a decoder of $\mathcal{R}^v$ with a total of S layers, let f_s be the feature map of the s-th layer, the s-th feature of the concatenation is computed as:

$$z_s = \begin{cases} \mathrm{Conv}_3\left(\mathrm{Concat}\left(\mathrm{Conv}_1(f_s), \mathrm{Up}(z_{s-1})\right)\right), & s \geq 2 \\ \mathrm{Conv}_1(f_s), & s = 1 \end{cases}. \tag{3}$$

where $\mathrm{Up}(\cdot)$ is an upsampling layer with a scaling factor of 2 to align the scales between two feature maps; $\mathrm{Conv}_1(\cdot)$ is a 1×1 convolutional layer to reduce channel dimensions; $\mathrm{Conv}_3(\cdot)$ is a 3×3 convolutional layer to suppress the aliasing effect of upsampling; $\mathrm{Concat}(\cdot)$ indicates the concatenation of two tensors. Finally, the last layer of the feature pyramid z_S is combined with the reconstructed image $\tilde{X}$ to get the enhanced feature F in the following way:

$$F = \mathrm{Concat}(\mathrm{Conv}_3(\tilde{X}), \mathrm{Conv}_3(z_S)) \tag{4}$$

Low-pass residual-guided attention. The distinguishable features are contained in the restored regional, color, and textural information of the image. Thus, it is possible to leverage the reconstruction residual to provide spatial attention to improve intra-view learning. However, one challenge is that, since the original image X is involved in computing the residual, both stable and unstable features in the original image potentially remain in the residual. As discussed earlier, unstable features that are detrimental to generalization and robustness should be avoided. Prior studies have found that these unstable features are low-level artifacts that mainly cluster in high-frequency components [Frank *et al.*, 2020; Durall *et al.*, 2020]. Thus, we propose only using the low-frequency residual to guide the classifier to focus on more stable features. Given an image X and its reconstructed version $\tilde{X}$, the low-frequency residual is:

$$M = |\mathcal{H}(X) - \mathcal{H}(\tilde{X})|, \tag{5}$$

where $\mathcal{H}(\cdot)$ is the first-order low-pass Butterworth filter and $|\cdot|$ is the absolute function. An attention mechanism is then devised to exploit the low-frequency residual. A functional network is used to process M to get the attention map, i.e., $\hat{M} = \mathcal{G}(M)$, where $\mathcal{G}(\cdot)$ consists of a 7×7 convolutional layer, an average pooling layer and a sigmoid function. The attention map is applied to the enhanced feature F in Eq. 4 to obtain the residual-guided feature:

$$\hat{F} = \hat{M} \otimes \mathrm{Conv}_3(F), \tag{6}$$

where $\otimes$ indicates the element-wise multiplication.

3.3 Inter-View Classification Learning

When the intra-view feature enhancement is ready, we can get a set of features $\{\hat{F}^v\}_{v=1}^N$ corresponding to different views. For each view, an independent neural network classifier $\mathcal{C}^v$ is trained on the feature $\hat{F}^v$. Since the features provide view-specific information, the classifiers will learn diverse representations and contribute differently facing the same data instance. To ensure the complementarity and interactivity across different views during training, we propose a self-adaptive cross-view loss fusion strategy.

Self-adaptive loss fusion. The self-adaptive loss fusion strategy aims to combine the losses of different classifiers using adaptive weights, such that the importance of each view-specific representation can be estimated and respected in the final decision. The weights are learned and autonomously adjusted during training. Formally, given a view-specific feature instance $\hat{F}^v$ and the corresponding label y ($y = 0$ if the the sample is a real image, otherwise 1), let p^v be the probability that the sample is fake predicted by $\mathcal{C}^v$. The training of $\mathcal{C}^v$ is supervised by minimizing the cross-entropy loss:

$$\min L_{ce}^v := -[y \log(p^v) + (1-y)\log(1-p^v)]. \tag{7}$$

The self-adaptive loss fusion strategy can be denoted as a minimization problem with respect to the weights $\boldsymbol{\beta}$:

$$\min_{\boldsymbol{\beta}} \sum_{v=1}^{N} \beta_v^\tau L_{ce}^v \quad s.t. \quad \boldsymbol{\beta}^\top \mathbf{1} = \mathbf{1}, \beta_v \geq 0, \tag{8}$$

where $\tau > 1$ is the power exponent parameter to avoid the trivial solution of $\boldsymbol{\beta}$ during the classification. In the inference stage, the decision is made on the average predicted probability of fake (p^v) over all classifiers, i.e., $p_{fake} = \frac{1}{3}\sum_v p^v$, with a threshold of 0.5.

3.4 Optimization

The components of MCCL that require optimization include the parameters of $\{\mathcal{R}^v\}_{v=1}^N$, $\{\mathcal{C}^v\}_{v=1}^N$ and several building blocks for intra-view learning (for simplicity, the latter two are denoted in together as $\{\mathcal{C}^v\}_{v=1}^N$), as well as the self-adaptive loss weights $\boldsymbol{\beta}$. The optimization is performed in the following alternative way:

Update network parameters. The completion and classification networks with respect to different views are updated independently in parallel. For the view v, $\mathcal{R}^v$ and $\mathcal{C}^v$ can be updated sequentially by optimizing the corresponding loss functions Eq. 2 and Eq. 7. During the optimization, the loss weights $\boldsymbol{\beta}$ are fixed.

Update loss weights $\boldsymbol{\beta}$. Next, we fix the parameters of $\{\mathcal{R}^v\}_{v=1}^N$ and $\{\mathcal{C}^v\}_{v=1}^N$, and update $\boldsymbol{\beta}$ by solving Eq. 8. To satisfy the constraints, the Lagrangian function of Eq. 8 is

$$\mathcal{L}(\boldsymbol{\beta}, \zeta) = \sum_{v=1}^{N} \beta_v^{\tau} L_{ce}^v - \zeta(\sum_{v=1}^{N} \beta_v - 1) \tag{9}$$

where ζ is the Lagrange multiplier. By derivation of Eq. 9 with respect to β_v and ζ, the optimal solution of Eq. 8 is:

$$\beta_v = (L_{ce}^v)^{\frac{1}{1-\tau}} / \sum_{n=1}^{N} (L_{ce}^n)^{\frac{1}{1-\tau}} \tag{10}$$

4 Experiments

4.1 Datasets

Real images. Our experiments are conducted on facial images, given that human faces are the main target of deepfakes. We choose the large-scale facial image dataset CelebA [Liu *et al.*, 2015] and its high-quality version CelebA-HQ [Karras *et al.*, 2018] to perform evaluations at different resolutions. The CelebA images have a resolution of 128 × 128 and the CelebA-HQ images have a resolution of 1024×1024.

GAN-generated images. A total of six popular GAN types are considered: ProGAN [Karras *et al.*, 2018], CramerGAN [Bellemare *et al.*, 2017], SNGAN [Miyato *et al.*, 2018], MMDGAN [Li *et al.*, 2017], StyleGAN [Karras *et al.*, 2019] and StyleGAN2 [Karras *et al.*, 2020]. In the low-resolution setting, we follow the setting in [Yu *et al.*, 2019], using the pre-trained ProGAN, CramerGAN, SNGAN, and MMDGAN models [1] to generate fake faces. All the four GANs are pre-trained with CelebA. In the high-resolution setting, we adopt the dataset released by [He *et al.*, 2021] [2], which includes images generated by ProGAN, StyleGAN, and StyleGAN2. Note that the ProGAN and StyleGAN are pre-trained with CelebA-HQ, while the StyleGAN2 with another facial image dataset FFHQ [Karras *et al.*, 2019]. Since FFHQ has a larger diversity in terms of facial attributes compared with CelebA-HQ, StyleGAN2 is included for cross-domain evaluation. Table 1 shows the details of the dataset setting.

128x128	CelebA	ProGAN	CramerGAN	SNGAN	MMDGAN
Training Test	60k 6k	60k 6k	– 6k	– 6k	– 6k
1024x1024	CelebA-HQ	ProGAN	StyleGAN	StyleGAN2*	
Training Test	25k 2.5k	25k 2.5k	25k 2.5k	– 2.5k	

* Pre-trained in a different real image dataset FFHQ

Table 1: The details of the experimental dataset setting.

4.2 Implementation Details

The restorers and classifiers are implemented based on U-Net [Ronneberger *et al.*, 2015] and Xception [Chollet, 2017], respectively. The U-Net we use has five skip connection blocks (i.e., $S = 5$), and their output feature maps are employed to build the feature concatenation. We train the whole framework with a batch size of 80 using the Adam optimizer [Kingma and Ba, 2015]. The initial learning rate is 1e-3, and we reduce it to half after every ten epochs. τ in Eq. 8, and λ in Eq. 2 are empirically set to 4 and 10, respectively. We also use random Gaussian noise, color jitter, and blurring for data augmentation on the restorer side.

	ProGAN		CramerGAN		SNGAN		MMDGAN	
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
GAN-FP	99.5	99.8	52.1	55.4	53.6	70.4	48.2	53.2
2d-DCT	98.9	99.1	70.2	67.1	61.9	73.5	56.0	73.1
DA	99.5	99.9	72.1	78.3	63.1	70.0	54.7	71.7
FLP	95.1	98.3	81.3	81.7	**83.6**	80.1	70.2	82.0
SRR	**100.**	**100.**	88.2	**95.1**	70.3	81.5	77.7	84.5
MCCL (Ours)	**100.**	**100.**	**91.1**	89.2	80.2	**83.3**	**85.4**	**86.1**

Table 2: The results of cross-GAN detection in the 128×128 setting. **Bold** indicates the best score in each column.

4.3 Baseline Detection Models

We compare MCCL with two normal detectors that exploit a CNN to extract and learn features directly: the image-domain detector using GAN fingerprints (GAN-FP) [Yu *et al.*, 2019] and the frequency-domain detector based on 2D DCT coefficients (2d-DCT) [Frank *et al.*, 2020]. We also compare three state-of-the-art generalized and robust detection methods: the data augmentation-based method (DA) [Wang *et al.*, 2020], the frequency-level perturbation (FLP) [Jeong *et al.*, 2022b], and the super-resolution re-synthesis (SRR) [He *et al.*, 2021], each of which has specific strategies to improve robustness. The detection performance is evaluated by the classification accuracy (Acc.) and the average precision score (A.P.) commonly used in related studies [Wang *et al.*, 2020; Jeong *et al.*, 2022b].

4.4 Results of Generalization

Low-resolution setting. We train the detection model with the CelebA images and the corresponding 128×128 ProGAN images and test it with the ProGAN, CramerGAN, SNGAN, and MMDGAN images to evaluate the cross-GAN generalization ability. The results compared to five baselines are listed in Table 2. We conclude that: 1) The normal detectors, GAN-FP and 2d-DCT, are highly accurate for within-distribution detection but generalize poorly for cross-GAN detection, implying the risk of overfitting unstable features; 2) The other four detectors with specific strategies all get the cross-GAN detection performance improved. Our method MCCL achieves the best or second-best results spanning over all GANs, thanks to the novel multi-view representation enriching more robust, frequency-irrelevant features.

High-resolution setting. We follow the setting in [He *et al.*, 2021]: we train a ProGAN detector and a StyleGAN detector independently using CelebA-HQ and the corresponding GAN images, and test both on ProGAN, StyleGAN, and StyleGAN2 test samples to evaluate the within-distribution, cross-GAN and cross-domain performances, respectively. The results are summarized in Table 3. All detectors perform well

[1] https://github.com/ningyu1991/GANFingerprints

[2] https://github.com/SSAW14/BeyondtheSpectrum

	Within-distribution				Cross-GAN				Cross-GAN & Cross-domain			
	P→P		S→S		P→S		S→P		P→S2		S→S2	
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
GAN-FP	99.9	99.9	99.4	99.6	51.1	71.0	49.3	68.8	44.3	47.6	48.0	46.9
2d-DCT	99.9	99.9	99.8	99.9	90.1	91.5	93.0	92.1	62.1	60.0	93.8	90.2
DA	97.6	96.6	98.3	97.8	73.2	87.7	78.1	73.1	66.1	79.1	80.7	84.4
FLP	98.9	99.0	99.1	98.9	95.0	97.1	94.3	86.3	80.8	88.0	92.4	93.1
SRR	**100.**	**100.**	99.9	99.9	**99.1**	**99.4**	98.2	98.1	88.2	80.3	91.5	91.1
MCCL (Ours)	**100.**	**100.**	**100.**	**100.**	98.1	95.5	**99.2**	**98.9**	**95.3**	**90.0**	**97.7**	**96.0**

Table 3: The results of cross-GAN detection in the 1024 × 1024 setting. **Bold** indicates the best-in-column. P, S, S2 are short for ProGAN, StyleGAN and StyleGAN2, respectively. The right and left sides of → indicate the training and test sets, respectively.

	Clean		Blurring		Cropping		Compression		Noise		Mix		FGSM		PGD		SDN	
	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.
GAN-FP	99.5	99.8	49.6	67.4	44.9	77.5	8.7	45.8	9.0	49.1	19.3	66.6	11.1	15.5	8.1	22.1	13.4	45.0
2d-DCT	98.9	99.1	60.4	77.7	80.5	76.1	67.4	80.2	46.7	74.3	61.3	61.8	34.0	45.3	23.1	41.3	21.8	56.1
DA	99.5	99.9	83.2	**98.9**	51.8	64.1	84.0	**97.3**	74.3	80.2	85.5	91.0	43.4	66.7	40.1	54.4	56.7	67.0
FLP	95.1	98.3	96.1	90.2	71.6	77.0	80.3	74.3	90.9	91.1	84.7	89.9	56.1	60.7	49.4	67.0	43.2	60.1
SRR	**100.**	**100.**	92.1	93.0	97.9	96.1	90.7	93.3	92.0	88.8	89.6	90.6	67.1	75.2	64.8	77.1	87.2	91.1
MCCL (Ours)	**100.**	**100.**	**96.4**	98.5	**98.2**	**99.1**	**93.8**	96.9	**94.7**	**94.4**	**91.3**	**94.4**	**81.6**	**80.3**	**81.3**	**81.9**	**93.2**	**95.6**

Table 4: The results of robustness against 8 perturbation methods. **Bold** indicates the best score in each column.

in within-distribution detection. In the cross-GAN group, 2d-DCT becomes more robust compared with its low-resolution performance. The reason may be that with the resolution increasing, stable low-frequency features are naturally enriched, which can be easier captured by the classifier trained directly on spectra. We can also see that FLP, SRR, and MCCL improve more significantly than DA in this group because they reduce classifiers' dependency on unstable frequency features in a learnable way. Regarding the cross-GAN & cross-domain group, which is the most challenging, our method remarkably outperforms all baseline methods in both sub-groups, indicating great applicability to difficult detection scenarios.

4.5 Results of Robustness

We evaluate the robustness against perturbations using the 128 × 128 ProGAN detectors. We train detectors with the CelebA and ProGAN images, and test them with perturbed ProGAN samples. Unlike prior work mainly concerning common image manipulations [Frank *et al.*, 2020; Yu *et al.*, 2019; Wang *et al.*, 2020], we investigate a broader range of perturbations: (1) Common manipulations, including Blurring, Cropping, Compression, Noising and a mix of all. We follow the setting in [Frank *et al.*, 2020] to created the perturbations; (2) Adversarial perturbations including FGSM [Goodfellow *et al.*, 2014b] and PGD [Madry *et al.*, 2018]. The adversarial examples are crafted based on a vanilla Xception detector with the noise amount $\epsilon = 8/255$; and (3) Spectrum Difference Normalization (SDN) [Dong *et al.*, 2022], an attack specific to GAN-generated images that calibrates the spectra of fake images according to real images.

An example of samples modified by different perturbations is shown in Figure 4. Table 4 shows the results. Since most perturbations significantly modify the original frequency distribution of fake samples, the performance of normal detectors degrades rapidly, while the other four are relatively more resistant given the suppression of frequency overfitting. Among the four robust methods, our method achieves the best results regarding all perturbations except for the A.P. scores for blurring and compression. It is worth noting that, when confronted with much more challenging perturbations such as the adversarial attacks FGSM and PGD and the specific attack SDN, our method obtains Acc. and A.P. scores that are notably higher than other baselines.

Summary. We here discuss the superiority of MCCL over the other three robust detection methods DA, FLP and SRR. DA and FLP augment the frequency distributions in the training set by image processing or adversarial perturbation to improve the generalization to unknown frequency patterns. However, only augmentation is insufficient as it cannot cover all GAN-specific frequency patterns. Unlike the frequency-level augmentation, MCCL seeks to eliminate the frequency distributional gaps between real and fake images via image completion, which can enforce the subsequent classifiers to learn more stable, frequency-irrelevant features. SRR employs a super-resolution process to model real image distribution, which can be regarded as a single-view image completion task. MCCL outperforms SRR because of the multi-view representation and several novel learning strategies, which allow stable and robust feature representation from diverse types of information.

4.6 Analysis and Discussion

Ablation study. Two ablation studies are performed based on the 128 × 128 ProGAN detectors to show the significance of individual views and the effects of different modules . We report the average Acc. (mAcc.) and A.P. (mA.P.) scores for cross-model and cross-perturbation performance, as shown in Table 5. Regarding different view settings, the gray view leads to a relatively weaker robustness than the other two views, indicating that color information is less distinguishable than regional consistency and texture. The combination

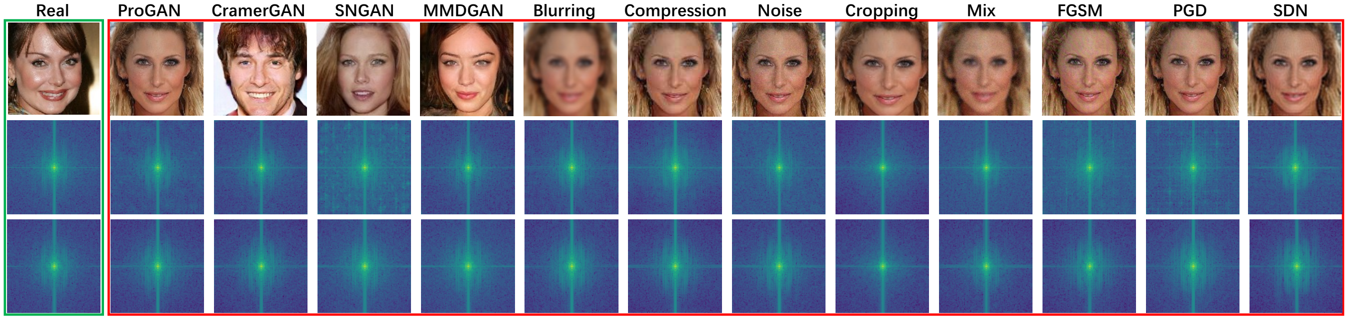

Figure 4: The visualization of real and different GAN-generated and perturbed fake samples (the 1st row) and the average FFT spectra before (the 2nd row) and after (the 3rd row) the Edge-to-RGB completion.

			Within-distribution		Cross-GAN		Cross-perturbation	
			Acc.	A.P.	mAcc.	mA.P.	mAcc.	mA.P.
Masked	Gray	Edge						
✓			99.6	99.0	82.1	81.3	83.3	87.5
	✓		99.1	99.3	67.2	72.1	71.1	77.0
		✓	100.0	100.0	78.9	83.4	88.8	90.1
MFC	LRA	ALF						
			93.7	97.0	75.2	73.1	81.4	74.9
✓			95.1	97.3	79.1	73.0	86.8	79.9
✓	✓		97.5	98.9	81.2	76.1	90.5	82.3
Final	version		100.0	100.0	85.6	86.2	91.3	92.6

Table 5: The results of ablation studies with different views or modules. MFC: Multi-scale Feature Concatenation. LRA: Low-pass Residual-guided Attention. ALF: Adaptive Loss Fusion.

of all views (i.e., the final version in Table 5) outperforms all individual views, confirming the effectiveness of multi-view representation and the capability of MCCL to capture and fuse different types of view-specific features for robust detection. Additionally, we emphasize that MCCL is fully flexible and extensible in view configuration, which can incorporate more quantities and types of views to enable stronger feature representations. Regarding different module settings, with more modules activated, the learning capacity of the model improves, enabling more effective feature representation.

Frequency analysis. One advantage of multi-view completion representation is that it helps reduce the classifier's reliance on unstable frequency patterns by aligning the frequency distributions between real and fake images. It works because the unstable frequency artifacts of fake samples are prior removed in the incomplete views, and then the missing frequency pattern is reconstructed and calibrated according to real images by the restorer pre-trained with real images. We provide a spectral analysis to confirm the frequency alignment by estimating the spectral distributions using the azimuthal integration over each radial frequency of the center-shifted FFT spectrum [Durall *et al.*, 2020]. Figure 5 shows the distribution curves of real images and images generated by different GANs before and after completion. The original distributions differ significantly between real and fake images and between different GANs. As a result, a CNN detector may easily overfit one specific frequency pattern that may not generalize to another. After the view-to-image completions, the spectral distributional gaps have become much closer. The alignment is more thorough in the edge-to-RGB completion than in the other two because edge sketches remove far more information from the original image than masked and gray views. These frequency-aligned training samples will force the classifier to focus on more stable, general, and frequency-insensitive features. Figure 4 provides a visualization of the averaged FFT spectra of different fake samples before and after the edge-to-RGB completion. The differences between real images and all types of fake images become visibly smaller after completion.

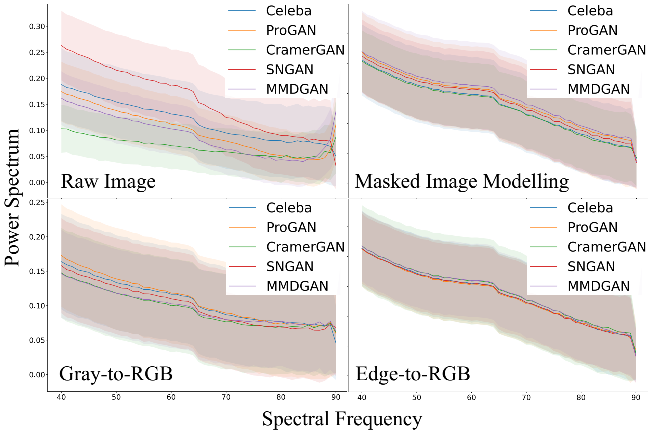

Figure 5: The spectral distributions of real images and fake images generated by different GANs before and after completion.

5 Conclusion

To overcome the generalization and robustness issues in GAN-generated image detection, we propose a novel framework incorporating multi-view completion learning and cross-view classification learning for a robust feature representation. Numerous experiments with varying cross-resolution, cross-GAN, and cross-perturbation settings validate the outperforming generalization and robustness of our proposed framework compared with the current state-of-the-art detectors. We also confirm the effect of reducing frequency reliance in deepfake detection, offering a potential route for future designs of robust deepfake detection.

Acknowledgements

This work is supported by an ARC Discovery Project DP230100246 from the Australian Research Council Australia.

References

[Barni *et al.*, 2020] Mauro Barni, Kassem Kallas, Ehsan Nowroozi, and Benedetta Tondi. Cnn detection of gan-generated face images based on cross-band co-occurrences analysis. In *2020 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2020.

[Bellemare *et al.*, 2017] Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017.

[Chollet, 2017] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[Dong *et al.*, 2022] Chengdong Dong, Ajay Kumar, and Eryun Liu. Think twice before detecting gan-generated fake images from their spectral domain imprints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7865–7874, 2022.

[Durall *et al.*, 2020] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7890–7899, 2020.

[Dzanic *et al.*, 2020] Tarik Dzanic, Karan Shah, and Freddie Witherden. Fourier spectrum discrepancies in deep network generated images. *Advances in neural information processing systems*, 33:3022–3032, 2020.

[Frank *et al.*, 2020] 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.

[Goodfellow *et al.*, 2014a] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[Goodfellow *et al.*, 2014b] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[He *et al.*, 2021] Yang He, Ning Yu, Margret Keuper, and Mario Fritz. Beyond the spectrum: Detecting deepfakes via re-synthesis. In *Thirtieth International Joint Conference on Artificial Intelligence*, pages 2534–2541. IJCAI, 2021.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[Hu *et al.*, 2021] Shu Hu, Yuezun Li, and Siwei Lyu. Exposing gan-generated faces using inconsistent corneal specular highlights. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2500–2504. IEEE, 2021.

[Huang *et al.*, 2020] Yihao Huang, Felix Juefei-Xu, Run Wang, Qing Guo, Lei Ma, Xiaofei Xie, Jianwen Li, Weikai Miao, Yang Liu, and Geguang Pu. Fakepolisher: Making deepfakes more detection-evasive by shallow reconstruction. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1217–1226, 2020.

[Jeon *et al.*, 2020] Hyeonseong Jeon, Youngoh Bang, and Simon S Woo. Fdftnet: Facing off fake images using fake detection fine-tuning network. In *IFIP international conference on ICT systems security and privacy protection*, pages 416–430. Springer, 2020.

[Jeong *et al.*, 2022a] Yonghyun Jeong, Doyeon Kim, Seungjai Min, Seongho Joe, Youngjune Gwon, and Jongwon Choi. Bihpf: Bilateral high-pass filters for robust deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 48–57, 2022.

[Jeong *et al.*, 2022b] Yonghyun Jeong, Doyeon Kim, Youngmin Ro, and Jongwon Choi. Frepgan: Robust deepfake detection using frequency-level perturbations. In *Proceedings of the AAAI conference on artificial intelligence*, 2022.

[Jung and Keuper, 2021] Steffen Jung and Margret Keuper. Spectral distribution aware image generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1734–1742, 2021.

[Karras *et al.*, 2018] 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.

[Karras *et al.*, 2019] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[Karras *et al.*, 2020] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

[Li *et al.*, 2017] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30, 2017.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[Liu *et al.*, 2020] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8060–8069, 2020.

[Liu *et al.*, 2021] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 772–781, 2021.

[Liu *et al.*, 2022] Chi Liu, Huajie Chen, Tianqing Zhu, Jun Zhang, and Wanlei Zhou. Making deepfakes more spurious: evading deep face forgery detection via trace removal attack. *arXiv preprint arXiv:2203.11433*, 2022.

[Madry *et al.*, 2018] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[Marra *et al.*, 2018] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of gan-generated fake images over social networks. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 384–389. IEEE, 2018.

[Marra *et al.*, 2019a] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 506–511. IEEE, 2019.

[Marra *et al.*, 2019b] Francesco Marra, Cristiano Saltori, Giulia Boato, and Luisa Verdoliva. Incremental learning for the detection and classification of gan-generated images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2019.

[McCloskey and Albright, 2019] Scott McCloskey and Michael Albright. Detecting gan-generated imagery using saturation cues. In *2019 IEEE international conference on image processing (ICIP)*, pages 4584–4588. IEEE, 2019.

[Mi *et al.*, 2020] Zhongjie Mi, Xinghao Jiang, Tanfeng Sun, and Ke Xu. Gan-generated image detection with self-attention mechanism against gan generator defect. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):969–981, 2020.

[Miyato *et al.*, 2018] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[Nataraj *et al.*, 2019] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, BS Manjunath, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, and Amit K Roy-Chowdhury. Detecting gan generated fake images using co-occurrence matrices. *Electronic Imaging*, 2019(5):532–1, 2019.

[Qian *et al.*, 2020] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer, 2020.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[Ruff *et al.*, 2020] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*, 2020.

[Tariq *et al.*, 2018] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S Woo. Detecting both machine and human created fake face images in the wild. In *Proceedings of the 2nd international workshop on multimedia privacy and security*, pages 81–87, 2018.

[Wang *et al.*, 2020] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020.

[Wang *et al.*, 2022] Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. M2tr: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 615–623, 2022.

[Yu *et al.*, 2019] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7556–7566, 2019.

[Zhang *et al.*, 2019] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2019.