Forgery-aware Adaptive Transformer for Generalizable Synthetic Image Detection

Huan Liu[1,3*]   Zichang Tan[2]   Chuangchuang Tan[1,3]   Yunchao Wei[1,3]   Yao Zhao[1,3]   Jingdong Wang[2]

[1]Institute of Information Science, Beijing Jiaotong University   [2]Baidu VIS

[3]Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China

{liu.huan,tanchuangchuang,yunchao.wei,yzhao}@bjtu.edu.cn {tanzichang,wangjingdong}@baidu.com

Abstract

In this paper, we study the problem of generalizable synthetic image detection, aiming to detect forgery images from diverse generative methods, e.g., GANs and diffusion models. Cutting-edge solutions start to explore the benefits of pre-trained models, and mainly follow the fixed paradigm of solely training an attached classifier, e.g., combining frozen CLIP-ViT with a learnable linear layer in UniFD [35]. However, our analysis shows that such a fixed paradigm is prone to yield detectors with insufficient learning regarding forgery representations. We attribute the key challenge to the lack of forgery adaptation, and present a novel forgery-aware adaptive transformer approach, namely FatFormer. Based on the pre-trained vision-language spaces of CLIP, FatFormer introduces two core designs for the adaption to build generalized forgery representations. First, motivated by the fact that both image and frequency analysis are essential for synthetic image detection, we develop a forgery-aware adapter to adapt image features to discern and integrate local forgery traces within image and frequency domains. Second, we find that considering the contrastive objectives between adapted image features and text prompt embeddings, a previously overlooked aspect, results in a nontrivial generalization improvement. Accordingly, we introduce language-guided alignment to supervise the forgery adaptation with image and text prompts in FatFormer. Experiments show that, by coupling these two designs, our approach tuned on 4-class ProGAN data attains a remarkable detection performance, achieving an average of 98% accuracy to unseen GANs, and surprisingly generalizes to unseen diffusion models with 95% accuracy.

1. Introduction

Recent years have witnessed the emergence and advancement of generative models, such as GANs [13, 23–25] and

---

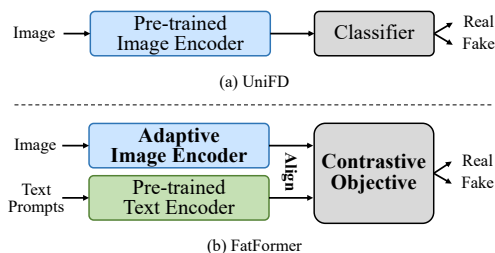*Work done when H. Liu is a long-term intern at Baidu.

Figure 1. **Comparison with fixed pre-trained paradigm.** Here, we illustrate the overview of UniFD [35] and our FatFormer. In contrast to training an attached classifier, FatFormer builds a forgery-aware adaptive transformer by aligning the representations of image and text prompts via contrastive objectives.

diffusion models [9, 14, 16, 34]. These models enable the creation of hyper-realistic synthetic images, thus raising the wide concerns of potential abuse and privacy threats. In response to such security issues, various forgery detection methods [11, 20, 21, 44, 45] have been developed, *e.g.*, image-based methods [3, 48] focusing on low-level visual artifacts and frequency-based methods [12, 37] relying on high-frequency pattern analysis. However, we observe big performance degradation when applying them to unseen images created by GANs or more recent diffusion models. How to address this problem has seen significant interest.

Recent approaches [35, 46] turn to explore the utilization of pre-trained models, following the fixed pre-trained paradigm of solely training an attached classifier, as shown in Figure 1 (a). A notable example in this field is the UniFD proposed by Ojha *et al.* [35], where a pre-trained CLIP-ViT [10, 38] is employed to encode images into image features without learning. Subsequently, a linear layer is tuned as a classifier to determine the credibility of inputs. At a very high level, their key to success is the employment of a pre-trained model in a frozen state, thus providing a learned universal representation (from the pre-training), yet not explicitly tuned in the current synthetic image detection task. In this way, such a representation will never be overfitted during training and thus preserves reasonable generalizability.
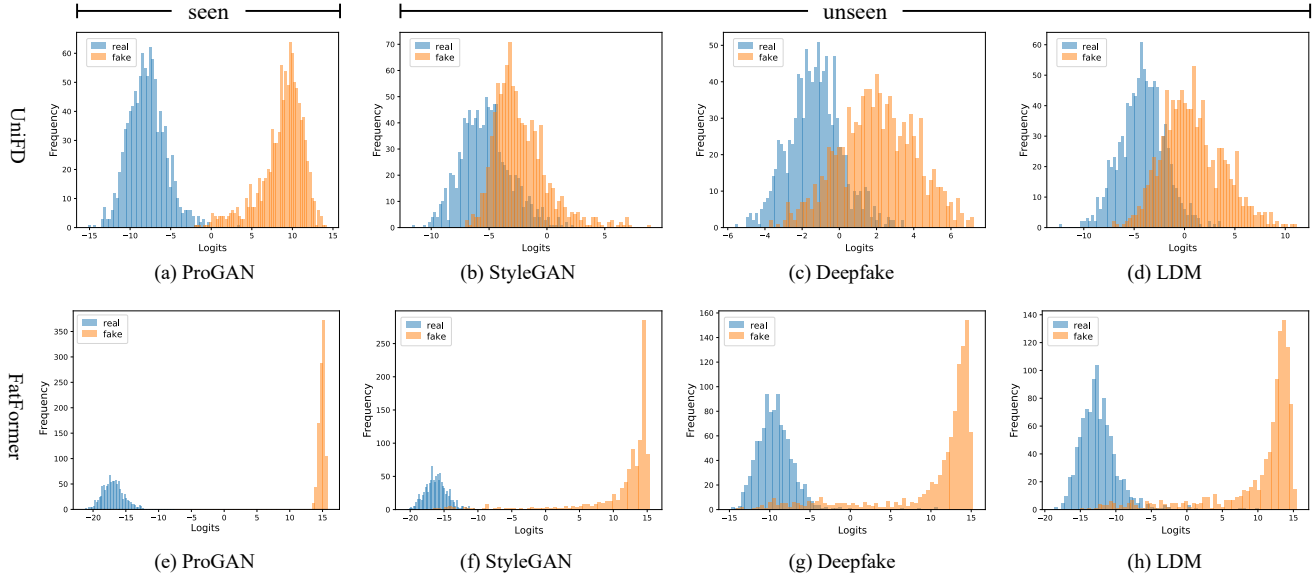
Figure 2. **Logit distributions of extracted forgery features.** We compare the state-of-the-art UniFD [35] and our FatFormer with forgery adaptation, both tuned with 4-class ProGAN [23] data. A total of four testing GANs and diffusion models are considered, including ProGAN [23], StyleGAN [24], Deepfake [42] and LDM [41], each randomly sampled 1k real and 1k fake images. Best view in color.

However, we consider that such a frozen operation adopted by UniFD will also limit the capability of pre-trained models for learning strong and pertinent forgery features.

To verify our assumption, we qualitatively study the forgery discrimination of the fixed pre-trained paradigm by visualizing the logit distributions of UniFD [35] across various generative models, as depicted in the top row of Figure 2. The distribution reflects the degree of separation between 'real' and 'fake' during testing, thereby offering the extent of generalization of extracted forgery representations. One can see that there is a large overlap of 'real' and 'fake' regions when facing unseen GANs or diffusion models (Figure 2 (b)-(d)), mistakenly, to identify these forgeries as 'real' class. Moreover, even in the case of Pro-GAN [23] testing samples, which employ the identical generative model as the training data, the distinction between 'read' and 'fake' elements becomes increasingly indistinct (Figure 2 (a) *vs.* (e)). We conclude that the fixed pre-trained paradigm is prone to yield detectors with insufficient learning regarding forgery artifacts, and attribute the key challenge to the lack of forgery adaptation that limits the full unleashing of potentials embedded in pre-trained models.

Driven by this analysis, we present a novel **F**orgery-aware **a**daptive **t**rans**Former** approach (Figure 1 (b)), named FatFormer, for generalizable synthetic image detection. In alignment with UniFD [35], FatFormer investigates CLIP [38] as the pre-trained model, which consists of a ViT [10] image encoder and a transformer [47] text encoder. Based on the pre-trained vision-language spaces of CLIP, our approach achieves the forgery adaptation by incorporat-

ing two core designs, ultimately obtaining well-generalized forgery representations with a distinct boundary between real and fake classes (Figure 2 (e)-(h)).

First, motivated by the fact that both image and frequency domains are important for synthetic image detection, a forgery-aware adapter (FAA) is developed, comprising a pair of image and frequency forgery extractors. In the image domain, a lightweight convolution module is employed for extracting low-level forgery artifacts, such as blur textures and color mismatch [29]. On the other hand, for the frequency domain, we construct a grouped attention mechanism that dynamically aggregates frequency clues from different frequency bands of discrete wavelet transform (DWT) [32]. By integrating these diverse forgery traces, FAA builds a comprehensive local viewpoint of image features essential for effective forgery adaptation.

Second, instead of utilizing the binary cross-entropy loss applied to image features, we consider the contrastive objectives between image and text prompts, a previously overlooked aspect. This novel direction is inspired by the natural language supervision in CLIP-ViT's pre-training, typically more robust to overfitting by optimizing the similarity between image features and text prompt embeddings [18]. Accordingly, language-guided alignment (LGA) is proposed, which encompasses a patch-based enhancer, designed to enrich the contextual relevance of text prompts by conditioning them on image patch tokens, as well as a text-guided interactor, that serves to align local image patch tokens with global text prompt embeddings, thereby directing the image encoder to concentrate on forgery-related representa-
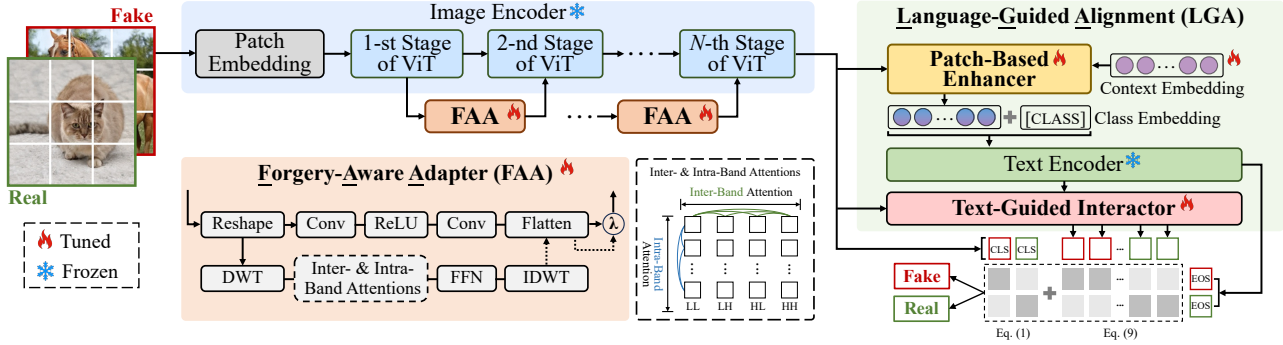
Figure 3. **Our FatFormer architecture.** The ViT image encoder integrates forgery-aware adapters to effectively extract visual forgery features from input images. To supervise the forgery adaptation process, language-guided alignment is introduced. Specifically, taking two input images for example, we maximize the cosine similarities between paired (dark gray squares) image features and text prompt embeddings, while minimizing the unpaired ones (light gray squares). In testing, only the test image is required to calculate the forgery probability via a softmax of these similarities. Squared 'CLS' and 'EOS' represent the image CLS tokens and text prompt embeddings.

tions. Empirical results show that the forgery adaptation supervised by LGA obtains more generalized forgery representations, thus improving the generalizability of synthetic image detection.

Our adaptive approach FatFormer significantly outperforms recent methods with the fixed pre-trained paradigm. Notably, we achieve 98.4% ACC and 99.7% AP on 8 types of GANs, and 95.0% ACC and 98.8% AP on 10 types of unseen diffusion images, using limited ProGAN training data. We hope our findings can facilitate the development of pre-trained paradigms in this field.

2. Related Work

Synthetic image detecting. Due to the increasing concerns about generative models, many works are proposed to address the problem of synthetic image detection, which can be roughly divided into image-based methods [33, 42, 49, 52], frequency-based methods [12, 20, 21], and pre-trained-based methods [35, 46]. For instance, Yu *et al.* [50] find images generated by GANs have unique fingerprints, which can be utilized as forgery traces for detection. Wang *et al.* [48] adopt various data augmentations and large-scale GAN images to improve the generalization to unseen testing data. Qian *et al.* [37] introduce frequency analysis into the detection framework, using local frequency statistics and decomposed high-frequency components for forgery detection. More recently, many works have focused on the fixed pre-trained paradigm of freezing the pre-trained model and adopting an attached classifier for forgery detection. For example, Lgrad [46] turns the detection problem into a pre-trained-based transformation-dependent problem, and utilizes gradient features from the frozen pre-trained model as forgery cues. Furthermore, Ojha *et al.* [35] propose UniFD to explore the potential of the vision-language model, *i.e.*, CLIP [38], for synthetic image detection. They observe that

training a deep network fails to detect fake images from new breeds and employs the frozen CLIP-ViT [10, 38] to extract forgery features, followed by a linear classifier.

In this paper, our motivation is different from the closely-related approach UniFD [35]. UniFD attempts to adopt a frozen pre-trained model to extract forgery representations 'without learning'. In contrast, our approach aims to demonstrate that the forgery adaptation of pre-trained models is essential for the generalizability of synthetic image detection.

Efficient transfer learning. The latest progress in transfer learning shows the potential for efficient fine-tuning of pre-trained models, especially in the NLP field. Unlike traditional strategies, such as linear-probing [15] and full fine-tuning [56], efficient transfer learning only adds learnable modules with a few parameters, such as prompt learning [27] and adapter-based methods [17, 19]. Inspired by this, many efficient transfer learning works are proposed for vision [5, 22] and vision-language models [53, 54]. Unlike UniFD [35] with linear probing, this paper investigates the efficient transfer learning for generalizable synthetic image detection and first proposes an adaptive transformer with contrastive objectives.

3. FatFormer

3.1. Overview

The overall structure of FatFormer is illustrated in Figure 3. FatFormer is composed of two pre-trained encoders for both image and text prompts, as well as the proposed forgery-aware adapter (Section 3.2) and language-guided alignment (Section 3.3). This framework predicts the forgery probability by calculating the softmax of cosine similarities between image features and text prompt embeddings.

Vanilla CLIP. Following UniFD [35], we adopt CLIP [38] as the pre-trained model with a ViT [10] image encoder and

transformer [47] text encoder, respectively. Given an image $x \in \mathbb{R}^{3 \times H \times W}$, with height H and width W, CLIP converts it into a D-dimensional image features $f_{img} \in \mathbb{R}^{(1+N) \times D}$, where 1 represents the image CLS token, $N = \frac{HW}{P^2}$ denotes the image patch tokens and P is the patch size. Meanwhile, the text encoder takes language text t and generates the text prompt embeddings $f_{text} \in \mathbb{R}^{M \times D}$ from the appended EOS tokens in the text encoder, where M denotes the number of classes (in this paper, $M = 2$). Two encoders are jointly trained to optimize the cosine similarity between the image CLS token and text prompt embeddings using contrastive loss. After pre-training, we can utilize the re-assembled text descriptions for zero-shot testing, *e.g.*, a simple template of 'this photo is [CLASS]', where '[CLASS]' is replaced by class names like 'real' or 'fake'. Given the testing image and text prompts, we have the predicted similarity of class $i \in \{0, 1\}$, where 0 represents 'real' and 1 is 'fake', as follows

$$S(i) = \cos(f_{img}^{(0)}, f_{text}^{(i)}), \tag{1}$$

where $\cos(\cdot)$ is the cosine similarity, $f_{img}^{(0)}$ denotes the image CLS token at index 0 of f_{img}. Further, the corresponding possibility can be derived via a softmax function

$$P(i) = \frac{\exp(S(i)/\tau)}{\sum_k \exp(S(k)/\tau)}, \tag{2}$$

where τ is the temperature parameter.

3.2. Forgery-aware adapter (FAA)

To adapt the image features for effective forgery adaptation, we insert forgery-aware adapters to bridge adjacent ViT stages, each encompassing multiple ViT layers, in the image encoder, as shown in Figure 3. These adapters discern and integrate forgery traces within both image and frequency domains, enabling a comprehensive local viewpoint of image features.

Image forgery extractor. In the image domain, FAA constructs a lightweight image forgery extractor, comprising two convolution layers and a ReLU layer for capturing low-level image artifacts, as follows

$$\hat{g}_{img}^{(j)} = \text{Conv}(\text{ReLU}(\text{Conv}(g_{img}^{(j)}))), \tag{3}$$

where $\hat{g}_{img}^{(j)}$ represents the adapted forgery-aware image features from FAA in j-th ViT stage, and $g_{img}^{(j)}$ is the vanilla features from the last multi-head attention module in j-th ViT stage. Here, we omit the reshape operators.

Frequency forgery extractor. For the frequency domain, a grouped attention mechanism is proposed to mine forgery traces in the frequency bands of discrete wavelet transform (DWT) [32]. Although previous detection methods [21, 37]

adopt fast Fourier transform [1] and discrete cosine transform [40], they destroy the position information [28] in the transformed frequency domain, which is crucial in the context of attention modeling [10]. Thus, we utilize DWT as the transform function, retaining the spatial structure of image features, which decomposes the inputs into 4 distinct frequency bands, including LL, LH, HL, and HH. Here, combinations of 'L' and 'H' represent the combined low and high pass filters. Then, two grouped attention modules, *i.e.*, inter-band attention and intra-band attention, are proposed for the extraction of frequency clues. As indicated in Figure 3, the inter-band attention explicitly explores the interactions across diverse frequency bands, while the intra-band attention builds interactions within each frequency band. This design achieves the dynamical aggregation of different positions and bands, rather than manual weighting like F3Net [37]. In practice, we implement them with multi-head attention modules [47]. Finally, FFN and inverse discrete wavelet transform (IDWT) are used to obtain forgery-aware frequency features $\hat{g}_{freq}^{(j)}$, which are transformed back into the image domain for further incorporation.

To prevent introducing hyper-parameters, we leverage a learnable scale factor λ to control the information from image and frequency domains as the final adapted image features of j-th stage of ViT, which will be sent to the first multi-head attention module in the next $(j+1)$-th stage.

$$\hat{g}^{(j)} = \hat{g}_{img}^{(j)} + \lambda \cdot \hat{g}_{freq}^{(j)}. \tag{4}$$

3.3. Language-guided alignment (LGA)

To supervise the forgery adaptation of FatFormer, language-guided alignment is proposed by considering the contrastive objectives between image and text prompts. In a bit more detail, LGA has a patch-based enhancer that enriches the context of text prompts, and a text-guided interactor that aligns the local image patch tokens with global text prompt embeddings. Finally, we implement an augmented contrastive objective for the loss calculation.

Patch-based enhancer. Instead of using hand-crafted templates as prompts, FatFormer has a soft prompt design by adopting auto context embeddings, following [53, 54]. Since synthetic image detection relies on local forgery details [4, 51], we develop a patch-based enhancer to enhance the contextual relevance of prompts via the condition of local image patch tokens, deriving forgery-relevant prompts context. Specifically, we first compute the image patch tokens $f_{img}^{(1:N)} \in \mathbb{R}^{N \times D}$ in the image encoder. Then, given C context embeddings $p_{ctx} \in \mathbb{R}^{C \times D}$, we have

$$A_{pbe} = p_{ctx} \cdot (f_{img}^{(1:N)})^T, \tag{5}$$

where $A_{pbe} \in \mathbb{R}^{C \times N}$ is the similarity matrix in patch-based enhancer. We use A_{pbe} to represent the intensity of image

patch tokens for constructing each context embedding, as follows

$$\hat{p}_{ctx} = \text{softmax}(A_{pbe}) \cdot f_{img}^{(1:N)} + p_{ctx}. \qquad (6)$$

Finally, we can obtain the set of possible text prompts by combining the enhanced context $\hat{p}_{ctx}$ and M [CLASS] embeddings, and send them to the text encoder.

Text-guided interactor. To guide the image encoder focusing on forgery-related representation, we propose a text-guided interactor, which aligns the local image patch tokens with global text prompt embeddings. Specifically, given the text prompt embeddings f_{text} from text encoder and image patch tokens $f_{img}^{(1:N)}$, our text-guided interactor calculates the similarity A_{tgi} between them by

$$A_{tgi} = f_{img}^{(1:N)} \cdot (f_{text})^T. \qquad (7)$$

Similar to Eq. (6), with A_{tgi}, sized $\mathbb{R}^{N \times M}$, we align the image patch tokens with text prompt embeddings by adaptively augmenting text representations, as follows

$$\hat{f}_{img}^{(1:N)} = \text{softmax}(A_{tgi}) \cdot f_{text} + f_{img}^{(1:N)}, \qquad (8)$$

where $\hat{f}_{img}^{(1:N)}$ denotes the aligned image patch tokens. Together with the augmented contrastive objectives, the image encoder is guided to concentrate on forgery-related representation within each distinct image patch.

Augmented contrastive objectives. For the loss calculation, we consider augmented contrastive objectives that comprise two elements. The first is the cosine similarity in Eq. (1) same as the vanilla CLIP. The second is the similarity between text prompt embeddings and aligned image patch tokens $\hat{f}_{img}^{(1:N)}$. With $t \in [1, N]$ and $i \in \{0, 1\}$, we have

$$S'(i) = \frac{1}{N} \sum_t \cos(\hat{f}_{img}^{(t)}, f_{text}^{(i)}). \qquad (9)$$

By merging similarities from Eq. (1) and Eq. (9), our Fat-Former describes a augmented probability $\hat{P}(i)$ by a softmax function, as follows

$$\hat{P}(i) = \frac{\exp((S(i) + S'(i))/\tau)}{\sum_k \exp((S(k) + S'(k))/\tau)}. \qquad (10)$$

In practice, we apply the cross-entropy function on Eq. (10) with label $y \in \{0, 1\}$ to calculate contrastive loss like the origin CLIP, as follows

$$\mathcal{L} = -y \cdot \log \hat{P}(y) - (1 - y) \cdot \log(1 - \hat{P}(y)). \qquad (11)$$

4. Experiments

4.1. Settings

Datasets. As generative methods are always coming up, we follow the standard protocol [35, 46, 48] that limits the accessible training data to only one generative model, while testing on unseen data, such as synthetic images from other GANs and diffusion models. Specifically, we train FatFormer on the images generated by ProGAN [23] with two different settings, including 2-class (chair, horse) and 4-class (car, cat, chair, horse) data from [48]. For evaluation, we collect the testing GANs dataset provided in [48] and diffusion model datasets in [35, 49], which contain synthetic images and the corresponding real images. The testing GANs dataset includes ProGAN [23], StyleGAN [24], StyleGAN2 [25], BigGAN [2], CycleGAN [55], StarGAN [7], GauGAN [36] and DeepFake [42]. On the other hand, the diffusion part consists of PNDM [30], Guided [9], DALL-E [39], VQ-Diffusion [14], LDM [41], and Glide [34]. For LDM and Glide, we also consider their variants with different generating settings. More details can be found in their official papers.

Evaluation metric. The accuracy (ACC) and average precision (AP) are reported as the main metrics during evaluation for each generative model, following the standard process [35, 46, 48]. To better evaluate the overall model performance over the GANs and diffusion model datasets, we also adopt the mean of ACC and AP on each dataset, denoted as ACC_M and AP_M.

Implementation details. Our main training and testing settings follow the previous study [35]. The input images are first resized into 256×256, and then image cropping is adopted to derive the final resolution of 224×224. We apply random cropping and random horizontal flipping at training, while center cropping at testing, both with no other augmentations. The Adam [26] is utilized with betas of $(0.9, 0.999)$. We set the initial learning rate as 4×10^{-4}, training epochs as 25, and adopt a total batch size of 256. Besides, a learning rate schedule is used, decaying at every 10 epochs by a factor of 0.9.

4.2. Main results

This paper aims to build a better paradigm with pre-trained models for synthetic image detection. Therefore, we mainly compare our FatFormer with previous methods that adopt the fixed pre-trained paradigm, such as LGrad [46] and UniFD [35]. In addition, to show the effectiveness of our approach, we also consider comparisons with existing image-based [3, 45, 48] and frequency-based methods [11, 12, 20, 21, 37].

Comparisons on GANs dataset. Table 1 reports the comparisons on the GANs dataset [48] with two different training data settings. Results show that our FatFormer consistently exceeds pre-trained-based LGrad [46] and UniFD [35]. Specifically, under 4-class supervision, FatFormer outperforms the current state-of-the-art method UniFD by a significant 9.3% ACC and 1.4% AP with the same pre-trained CLIP model, achieving 98.4% ACC and 99.7%

Table 1. **Accuracy and average precision comparisons with state-of-the-art methods on GANs dataset.** We report the performance (in the formulation of ACC / AP) with two different training settings, including supervision from 2-class and 4-class ProGAN data, following [48]. Besides, we also provide the reference (Ref) for previous frameworks. † denotes only trained on self-blended images of FF++ [42]. The performance (ACC_M / AP_M) over the entire dataset is marked in gray . The **best results** are highlighted in **bold**.

	Methods	Ref	ProGAN	StyleGAN	StyleGAN2	BigGAN	CycleGAN	StarGAN	GauGAN	Deepfake	Mean
2-class supervision	Wang [48]	CVPR 2020	64.6 / 92.7	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	57.3 / 79.6
	Durall [11]	CVPR 2020	79.0 / 73.9	63.6 / 58.8	67.3 / 62.1	69.5 / 62.9	65.4 / 60.8	99.4 / 99.4	67.0 / 63.0	50.5 / 50.2	70.2 / 66.4
	Frank [12]	ICML 2020	85.7 / 81.3	73.1 / 68.5	75.0 / 70.9	76.9 / 70.8	86.5 / 80.8	85.0 / 77.0	67.3 / 65.3	50.1 / 55.3	75.0 / 71.2
	F3Net [37]	ECCV 2020	97.9 / 100.0	84.5 / 99.5	82.2 / 99.8	65.5 / 73.4	81.2 / 89.7	100.0 / 100.0	57.0 / 59.2	59.9 / 83.0	78.5 / 88.1
	BiHPF [20]	WACV 2022	87.4 / 87.4	71.6 / 74.1	77.0 / 81.1	82.6 / 80.6	86.0 / 86.6	93.8 / 80.8	75.3 / 88.2	53.7 / 54.0	78.4 / 79.1
	FrePGAN [21]	AAAI 2022	99.0 / 99.9	80.8 / 92.0	72.2 / 94.0	66.0 / 61.8	69.1 / 70.3	98.5 / 100.0	53.1 / 51.0	62.2 / 80.6	75.1 / 81.2
	LGrad [46]	CVPR 2023	99.8 / 100.0	**94.8 / 99.7**	**92.4 / 99.6**	82.5 / 92.4	85.9 / 94.7	99.7 / 99.9	73.7 / 83.2	60.6 / 67.8	86.2 / 92.2
	UniFD [35]	CVPR 2023	99.7 / 100.0	78.8 / 97.4	75.4 / 96.7	91.2 / 99.0	91.9 / 99.8	96.3 / 99.9	91.9 / 100.0	80.0 / 89.4	88.1 / 97.8
	Ours	−	**99.8 / 100.0**	87.7 / 97.4	91.1 / 99.3	**98.9 / 99.9**	**99.9 / 100.0**	**100.0 / 100.0**	**99.9 / 100.0**	**89.4 / 97.3**	**95.8 / 99.2**
4-class supervision	Wang [48]	CVPR 2020	91.4 / 99.4	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	67.1 / 86.9
	Durall [11]	CVPR 2020	81.1 / 74.4	54.4 / 52.6	66.8 / 62.0	60.1 / 56.3	69.0 / 64.0	98.1 / 98.1	61.9 / 57.4	50.2 / 50.0	67.7 / 64.4
	Frank [12]	ICML 2020	90.3 / 85.2	74.5 / 72.0	73.1 / 71.4	88.7 / 86.0	75.5 / 71.2	99.5 / 99.5	69.2 / 77.4	60.7 / 49.1	78.9 / 76.5
	PatchFor [3]	ECCV 2020	97.8 / 100.0	82.6 / 93.1	83.6 / 98.5	64.7 / 69.5	74.5 / 87.2	100.0 / 100.0	57.2 / 55.4	85.0 / 93.2	80.7 / 87.1
	F3Net [37]	ECCV 2020	99.4 / 100.0	92.6 / 99.7	88.0 / 99.8	65.3 / 69.9	76.4 / 84.3	100.0 / 100.0	58.1 / 56.7	63.5 / 78.8	80.4 / 86.2
	Blend† [45]	CVPR 2022	58.8 / 65.2	50.1 / 47.7	48.6 / 47.4	51.1 / 51.9	59.2 / 65.3	74.5 / 89.2	59.2 / 65.5	93.8 / 99.3	61.9 / 66.4
	BiHPF [20]	WACV 2022	90.7 / 86.2	76.9 / 75.1	76.2 / 74.7	84.9 / 81.7	81.9 / 78.9	94.4 / 94.4	69.5 / 78.1	54.4 / 54.6	78.6 / 77.9
	FrePGAN [21]	AAAI 2022	99.0 / 99.9	80.7 / 89.6	84.1 / 98.6	69.2 / 71.1	71.1 / 74.4	99.9 / 100.0	60.3 / 71.7	70.9 / 91.9	79.4 / 87.2
	LGrad [46]	CVPR 2023	99.9 / 100.0	94.8 / **99.9**	96.0 / 99.9	82.9 / 90.7	85.3 / 94.0	99.6 / 100.0	72.4 / 79.3	58.0 / 67.9	86.1 / 91.5
	UniFD [35]	CVPR 2023	99.7 / 100.0	89.0 / 98.7	83.9 / 98.4	90.5 / 99.1	87.9 / 99.8	91.4 / 100.0	89.9 / 100.0	80.2 / 90.2	89.1 / 98.3
	Ours	−	**99.9 / 100.0**	**97.2 / 99.8**	**98.8 / 99.9**	**99.5 / 100.0**	**99.3 / 100.0**	99.8 / 100.0	**99.4 / 100.0**	**93.2 / 98.0**	**98.4 / 99.7**

Table 2. **Accuracy and average precision comparisons with state-of-the-art methods on diffusion model dataset.** Models here are trained on the 4-class ProGAN data. We transpose the table for better readability. Notations are consistent with Table 1.

	Dataset	Wang [48]	Durall [11]	Frank [12]	PatchFor [3]	F3Net [37]	Blend† [45]	LGrad [46]	UniFD [35]	Ours
	PNDM	50.8 / 90.3	44.5 / 47.3	44.0 / 38.2	50.2 / 99.9	72.8 / 99.5	48.2 / 48.1	69.8 / 98.5	75.3 / 92.5	**99.3 / 100.0**
	Guided	54.9 / 66.6	40.6 / 42.3	53.4 / 52.5	74.2 / 81.4	69.2 / 70.8	58.3 / 63.4	**86.6 / 100.0**	75.7 / 85.1	76.1 / 92.0
	DALL-E	51.8 / 61.3	55.9 / 58.0	57.0 / 62.5	79.8 / 99.1	71.6 / 79.9	52.4 / 51.6	88.5 / 97.3	89.5 / 96.8	**98.8 / 99.8**
	VQ-Diffusion	50.0 / 71.0	38.6 / 38.3	51.7 / 66.7	100.0 / 100.0	100.0 / 100.0	77.1 / 82.6	96.3 / 100.0	83.5 / 97.7	**100.0 / 100.0**
LDM	200 steps	52.0 / 64.5	61.7 / 61.7	56.4 / 50.9	95.6 / **99.9**	73.4 / 83.3	52.6 / 51.9	94.2 / 99.1	90.2 / 97.1	**98.6** / 99.8
	200 w/ CFG	51.6 / 63.1	58.4 / 58.5	56.5 / 52.1	94.0 / **99.8**	80.7 / 89.1	51.9 / 52.6	**95.9** / 99.2	77.3 / 88.6	94.9 / 99.1
	100 steps	51.9 / 63.7	62.0 / 62.6	56.6 / 51.3	95.8 / 99.8	74.1 / 84.0	53.0 / 54.0	94.8 / 99.2	90.5 / 97.0	**98.7 / 99.9**
Glide	100-27	53.0 / 71.3	48.9 / 46.9	50.4 / 40.8	82.8 / 99.1	87.0 / 94.5	59.4 / 64.1	87.4 / 93.2	90.7 / 97.2	**94.4 / 99.1**
	50-27	54.2 / 76.0	51.7 / 49.9	52.0 / 42.3	84.9 / 98.8	88.5 / 95.4	64.2 / 68.3	90.7 / 95.1	91.1 / 97.4	**94.7 / 99.4**
	100-10	53.3 / 72.9	54.9 / 52.3	53.6 / 44.3	87.3 / **99.7**	88.3 / 95.4	58.8 / 63.2	89.4 / 94.9	90.1 / 97.0	**94.2** / 99.2
	Mean	52.4 / 70.1	51.7 / 51.8	53.2 / 50.2	84.5 / 97.8	80.6 / 89.2	57.6 / 60.0	89.4 / 97.7	85.4 / 94.6	**95.0 / 98.8**

AP. Besides, for the other 2-class supervision setting, similar trends are observed with the ones under 4-class supervision, when compared with pre-trained-based methods. Moreover, we also compare FatFormer with representative image-based [3, 45, 48] and frequency-based methods [11, 12, 20, 21, 37] in Table 1. Our approach can also easily outperform all of them with a larger improvement.

The above evidence indicates the necessity of forgery adaptation for pre-trained models. Beyond the impressive performance, more importantly, our FatFormer provides an effective paradigm of how to incorporate pre-trained models in the synthetic image detection task.

Comparisons on diffusion model dataset. To further demonstrate the effectiveness of FatFormer, we provide comparisons with existing detection methods on the diffusion model dataset [35]. The results are shown in Table 2. Note that all the compared methods are trained on 4-class ProGAN data. This test setting is more challenging as forged images are created by various diffusion models with completely different generating theories and processes from GANs. Surprisingly, FatFormer generalizes well for diffusion models, achieving 95.0% ACC and 98.8% AP.

Compared with pre-trained-based LGrad [46] and UniFD [35], FatFormer also works better than both of them when handling diffusion models. For example, our approach surpasses UniFD by 9.6% ACC and 4.2% AP. More-

Table 3. **Ablation experiments for FatFormer.** Evaluated on GANs dataset. Default settings are marked in gray.

(a) **Forgery-aware adapter implementations.** In the proposed framework, both image (img) and frequency (freq) domains are essential for building generalized forgery representation.

w/ img domain	w/ freq domain	ACC_M	AP_M
✓	✓	**98.4**	**99.7**
✓	×	95.4	99.6
×	✓	97.3	99.6

(b) **Frequency band interactions.** Both inter- and intra-band attentions are important for modeling forgery traces in the frequency domain.

interaction	ACC_M	AP_M
intra	97.4	99.7
inter	96.6	99.6
intra & inter	**98.4**	**99.7**

(c) **Benefits of supervision in vision-language space.** On the model only with img input, text is first added for building contrastive (contra) objectives. Then, we apply the proposed augmented (aug) contra strategy.

input modality	strategy	ACC_M	AP_M
only img	linear probe	95.3	99.2
img & text	contra	96.4	99.6
img & text	aug contra	**98.4**	**99.7**

(d) **Text prompt designs.** The auto embedding and img condition can benefit the performance, especially by considering the correlation between prompt and img patch tokens.

prompt designs	w/ img condition	ACC_M	AP_M
fixed template	×	95.5	99.6
auto embedding	×	96.4	99.6
auto embedding	CLS token	98.1	99.7
auto embedding	patch tokens	**98.4**	**99.7**

(e) **Model components.** Both components are essential in our FatFormer. We also conduct an extra experiment to test the zero-shot performance by removing the forgery-aware adapter and language-guided alignment.

module components	ACC_M	AP_M
none for zero-shot	66.6	74.3
forgery-aware adapter	95.3	99.2
language-guided alignment	91.5	98.1
forgery-aware adapter & language-guided alignment	**98.4**	**99.7**

over, we find that even with powerful CLIP as the pre-trained model, UniFD only achieves a similar result (about 85% ACC) like PatchFor [3]. We argue this is mainly because the fixed pre-trained paradigm is prone to yield detectors with insufficient learning regarding forgery artifacts. Thus, our FatFormer, which presents an adaptive transformer framework with forgery adaptation and reasonable contrastive objectives, can achieve much better results.

4.3. Ablation study

We conduct several ablation experiments to verify the effectiveness of key elements in our FatFormer. Unless specified, we report the mean of accuracy (ACC_M) and average precision (AP_M) on the GANs dataset under the training setting of 4-class ProGAN data.

Forgery-aware adapter implementations. We ablate the effects of considering the image domain and frequency domain in the forgery-aware adapter. The results are shown in Table 3a. We observe severe performance degradation when removing either of these two domains, especially for the frequency domain with over -3.0% ACC gaps. We conclude that both image and frequency domains are essential in FatFormer for synthetic image detection. The image forgery extractor collects the local low-level forgery artifacts, *e.g.*, blur textures, while the frequency forgery extractor explores and gathers the forgery clues among different frequency bands, together building a comprehensive local viewpoint for the adaptation of image features.

For the frequency forgery extractor, both interactions built by inter-band and intra-band attentions are important in our FatFormer. Table 3b shows the ablation.

Benefits of supervision in vision-language space. Table 3c provides the comparisons between different supervising strategies for FatFormer, including (i) linear probing with image modality, (ii) vanilla contrastive objectives between image CLS token and text prompt embeddings, which masked out the text-guided interactor, and (iii) our

augmented contrastive objectives. The results demonstrate that introducing text prompts for contrastive supervision benefits the generalization of detection. We conjecture this is mainly because CLIP provides a stable alignment between real image and text representation with pre-training, thus yielding a mismatching when handling a fake image with text prompts. As potential evidence, we find that only adopting LGA can still achieve an accuracy of 91.5% ACC (Table 3e). Besides, we observe that the proposed augmented contrastive objectives can further boost generalizability by directing the image encoder to concentrate on forgery-related representations, bringing a 2.0% ACC gain over the vanilla implementation.

Text prompt designs. Table 3d gives the results of constructing the text prompt with different prompt designs and image conditions. The results validate that both auto context embeddings and image conditions are important in text prompt designs. Compared with using a fixed hand-crafted template, *e.g.*, 'this photo is', the design of auto context embedding improves by 0.9% ACC, due to its abstract exploration in word embedding spaces. Besides, it is better to adopt image patch tokens as conditions to enhance these auto context embeddings, containing more local context details, rather than the global image CLS token.

Model components. Tabel 3e gives the ablation of two proposed model components, *i.e.*, forgery-aware adapter and language-guided alignment. Large performance drops (-6.9% ACC and -1.6% AP) are observed when adopting the previous fixed pre-trained paradigm by removing the forgery-aware adapter. This explains the necessity of forgery adaptation of pre-trained models. On the other hand, the proposed language-guided alignment, which considers the augmented contrastive objectives in the vision-language space, also provides better supervision for the forgery adaptation than simply adopting binary labels, bringing 3.1% ACC and 0.5% AP gains.
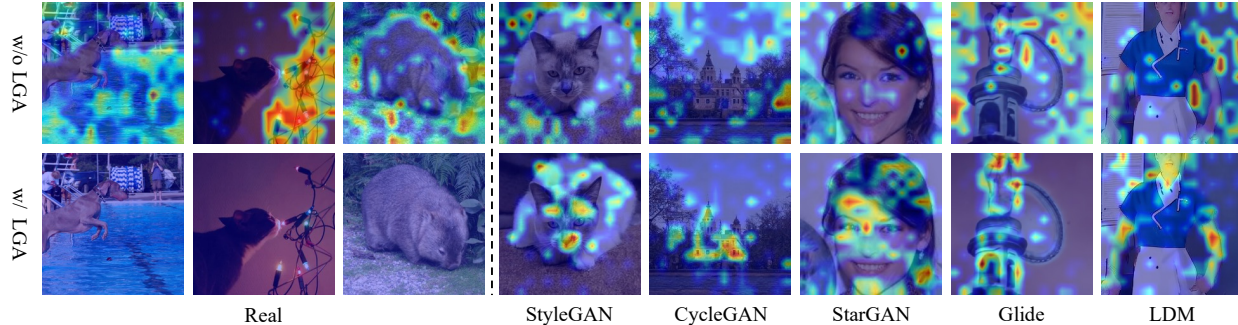
As shown in Figure 4, using language-guided align-

Figure 4. **Comparison of model attention with (w/) and without (w/o) language-guided alignment.** We visualize the gradient norm of FatFormer (second row) and FatFormer without language-guided alignment (first row) by [43]. FatFormer provides more responses among semantic foreground patches in fake images, while almost no response for real ones. The salient region is visualized by bright color.

ment obtains more concentration on semantic foreground patches, where anomalies, *e.g.*, unrealistic objects, textures, or structures often occur. Therefore, our FatFormer can obtain generalized forgery representations by focusing on local forgery details, resulting in the improvement of the generalizability of synthetic image detection.

4.4. More analysis

Here, we analyze our FatFormer on different architectures and pre-training strategies.

Analysis on different architectures. While FatFormer is constructed upon the identical CLIP framework [38] as employed in UniFD [35], the proposed forgery adaptation strategy is transferrable to alternative architectures. Presented in the upper section of Table 4 are the ACC_M and AP_M scores for four distinct architectures, including two variations of multi-modal structures pre-trained by CLIP and two variants of image-based Swin transformer [31] pre-trained on ImageNet 22k [8]. The comparisons between models with and without FatFormer verify the efficacy of integrating forgery adaptation among different pre-trained architectures, significantly facilitating the performance of synthetic image detection.

Analysis on different pre-training strategies. We further conduct an assessment of the efficacy of forgery adaptation across models employing different pre-training strategies. Utilizing ViT-L [10] as the baseline, we validate two well-known pre-training approaches: MAE [15] and CAE [6]. The evaluations are shown in the lower segment of Table 4. We observe that incorporating the forgery adaptation in our FatFormer can lead to a consistent increase in performance across diverse pre-training strategies, demonstrating the robustness and transferability of our approach.

5. Conclusion

In this paper, we present a novel adaptive transformer, Fat-Former, for generalizable synthetic image detection. With two core designs, including the forgery-aware adapter and

Table 4. **Analysis on different architectures and pre-training strategies.** Beyond UniFD, the forgery adaptation in FatFormer can also consistently boost various architectures and different pre-training strategies. We report the mean of ACC and AP (in the formulation of ACC_M / AP_M) on both GANs and diffusion model (DMs) datasets. 'IN-22K'= ImageNet 22k.

Architecture	Pre-training	w/ Ours	GANs	DMs
ViT-B/16 Text-512	CLIP [38]	×	83.8 / 94.4	77.2 / 91.1
		✓	**95.3 / 99.5**	**91.6 / 97.8**
ViT-L/14 Text-768	CLIP [38]	×	89.1 / 98.3	85.4 / 94.6
		✓	**98.4 / 99.7**	**95.0 / 98.8**
Swin-B	IN-22K [8]	×	82.5 / 93.8	72.2 / 88.8
		✓	**89.6 / 98.2**	**76.1 / 96.1**
Swin-L	IN-22K [8]	×	86.4 / 95.7	74.4 / 90.8
		✓	**90.7 / 98.4**	**79.3 / 96.7**
ViT-L/16	MAE [15]	×	75.7 / 92.8	70.9 / 92.3
		✓	**85.2 / 96.7**	**88.5 / 98.4**
	CAE [6]	×	76.1 / 95.9	64.9 / 91.7
		✓	**88.1 / 98.0**	**76.1 / 96.2**

language-guided alignment, for the forgery adaption of pre-trained models, the proposed approach outperforms the previous fixed pre-trained paradigm by a large margin. Besides, the forgery adaption in FatFormer is also flexible, which can be applied in various pre-trained architectures with different pre-training strategies. We hope FatFormer can provide insights for exploring better utilization of pre-trained models in the synthetic image detection field.

Limitations and future works. FatFormer generalizes well on most generative methods, while we still have space to improve in diffusion models, *e.g.*, Guided [9]. Elucidating the distinctions and associations among images produced by diffusion models and GANs is needed to build stronger forgery detectors. The investigation of this problem is left in future work. Besides, how to construct a better pretext task special for synthetic image detection in pre-training is also worth a deeper study.

References

[1] E Oran Brigham and RE Morrow. The fast fourier transform. *IEEE Spectrum*, 4(12):63–70, 1967. 4

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018. 5

[3] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *Proceedings of the European Conference on Computer Vision*, pages 103–120. Springer, 2020. 1, 5, 6, 7

[4] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1081–1088, 2021. 4

[5] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. 3

[6] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, pages 1–16, 2023. 8

[7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. 5

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 8

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1, 5, 8

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 1, 2, 3, 4, 8

[11] 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. 1, 5, 6

[12] 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, 3, 5, 6

[13] 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. 1

[14] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 1, 5

[15] 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. 3, 8

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[17] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 3

[18] Dapeng Hu, Shipeng Yan, Qizhengqiu Lu, HONG Lanqing, Hailin Hu, Yifan Zhang, Zhenguo Li, Xinchao Wang, and Jiashi Feng. How well does self-supervised pre-training perform with streaming data? In *International Conference on Learning Representations*, 2021. 2

[19] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 3

[20] 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. 1, 3, 5, 6

[21] 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*, pages 1060–1068, 2022. 1, 3, 4, 5, 6

[22] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 3

[23] 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. 1, 2, 5

[24] 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. 2, 5

[25] 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. 1, 5

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[27] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. 3

[28] Qiufu Li, Linlin Shen, Sheng Guo, and Zhihui Lai. Wavelet integrated cnns for noise-robust image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7245–7254, 2020. 4

[29] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020. 2

[30] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, 2021. 5

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 8

[32] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989. 2, 4

[33] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *IEEE Conference on Multimedia Information Processing and Retrieval*, pages 506–511. IEEE, 2019. 3

[34] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 1, 5

[35] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023. 1, 2, 3, 5, 6, 8

[36] 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. 5

[37] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Proceedings of the European Conference on Computer Vision*, pages 86–103. Springer, 2020. 1, 3, 4, 5, 6

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 8

[39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 5

[40] K Ramamohan Rao and Ping Yip. *Discrete cosine transform: algorithms, advantages, applications*. Academic Press Professional, Inc., USA, 1990. 4

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 5

[42] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–11, 2019. 2, 3, 5, 6

[43] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 618–626, 2017. 8

[44] Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2023. 1

[45] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022. 1, 5, 6

[46] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12105–12114, 2023. 1, 3, 5, 6

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4

[48] 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. 1, 3, 5, 6

[49] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3, 5

[50] 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. 3

[51] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021. 4

[52] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15023–15033, 2021. 3

[53] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 3, 4

[54] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3, 4

[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2223–2232, 2017. 5

[56] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the Institute of Electrical and Electronics Engineers*, 109(1):43–76, 2021. 3

A. Appendix

In this appendix, we first discuss the *potential negative societal impacts* (refer to Section A.1) that may arise in practical scenarios. Then, an in-depth exploration of *ablation studies* (explicated in Section A.2) is presented, delineating the influence of hyper-parameters employed within our approach. Lastly, a comprehensive analysis is conducted to assess the efficacy of forgery adaptation in enhancing *robustness* (outlined in Section A.3) against image perturbations.

A.1. Broader impacts

The development of synthetic image detection tools, while aiming to combat misinformation, may lead to unintended consequences in content moderation. Legitimate content that exhibits characteristics similar to forgeries may be mistakenly flagged, impacting normal information (based on image modality) sharing. These issues need further research and consideration when deploying this work to practical applications for content moderation.

A.2. More Ablations

We provide more ablation studies on the hyper-parameters used in our FatFormer. The training and evaluating settings are the same as Section 4.3.

Number of auto context embeddings. FatFormer combines the enhanced context embeddings and [CLASS] embeddings to construct the set of possible text prompts. Here, we ablate the effects of how a pre-defined number of context embeddings in text prompts affects the performance in the following table:

#embeddings	ACC_M	AP_M
4	97.6	99.0
8	**98.4**	**99.7**
16	97.8	99.6

One can see that 8 auto context embeddings are good enough and achieve better results than 16 embeddings. Thus, we set the number as 8 by default in this paper.

Number of forgery-aware adapters. To achieve effective forgery adaptation, FatFormer develops the forgery-aware adapter and integrates it with the ViT image encoder. The number of inserted forgery-aware adapters is to be explored. The following table lists the relevant ablations:

#adapters	ACC_M	AP_M
2	97.2	99.6
3	**98.4**	**99.7**
4	96.5	99.7

We observe that inserting 3 forgery-aware adapters in the image encoder is able to achieve good performance. Therefore, we set 3 as the default number of the forgery-aware adapter in our FatFormer.

Kernel size of image forgery extractor. To capture low-level image artifacts, we introduce a lightweight image forgery extractor in the proposed forgery-aware adapter, including two convolutional layers and a ReLU. We also explore settings of the kernel size of convolutional layers, as follows:

kernel size	ACC_M	AP_M
1	**98.4**	**99.7**
3	96.4	99.7
5	95.6	99.6

We find that using 1×1 kernel yields superior results in constructing the image forgery extractor. We conjecture that this is mainly because the intermediate image patch tokens in ViT encode high-level semantic information of different image patches, which may not provide useful low-level similarity among adjacent positions like the ones in traditional convolutional networks. Thus, larger kernels, designed to fuse adjacent patch tokens, may introduce disturbance to the modeling process of ViT and damage the performance.

A.3. Robustness on image perturbation

To evaluate the effects of forgery adaptation in FatFormer on robustness, we apply several common image perturbations to the test images, following [12, 46]. Specifically, we adopt random cropping, Gaussian blurring, JPEG compression, and Gaussian noising, each with a probability of 50%. The detailed perturbation configures can be found in [12]. Based on the GANs dataset, we compare our FatFormer with UniFD [35] and LGrad [46], which adopts the fixed pre-trained paradigm. The results are shown in the following table:

Perturbation	Method	ACC_M	AP_M
Gaussian blurring	LGrad	78.5	83.2
	UniFD	78.1	93.0
	FatFormer	**90.7**	**98.1**
random cropping	LGrad	85.0	91.9
	UniFD	88.9	98.1
	FatFormer	**98.2**	**99.7**
JPEG compression	LGrad	69.5	81.2
	UniFD	88.4	97.7
	FatFormer	**95.9**	**99.2**
Gaussian noising	LGrad	69.1	79.4
	UniFD	82.6	93.9
	FatFormer	**88.0**	**96.5**

It can be observed that our approach exceeds UniFD by a larger margin, *e.g.*, over $+12.0\%$ facing Gaussian blurring. This is mainly because FatFormer obtains well-generalized forgery representations with the proposed forgery adaption, as analyzed in Section 4.3.

Moreover, we also consider a more real-world scenario by combining all four types of perturbation. The results are illustrated in Figure 5. Compared with UniFD, our FatFormer also beats it on all testing GAN methods, further suggesting the robustness improvement brought by forgery adaptation.
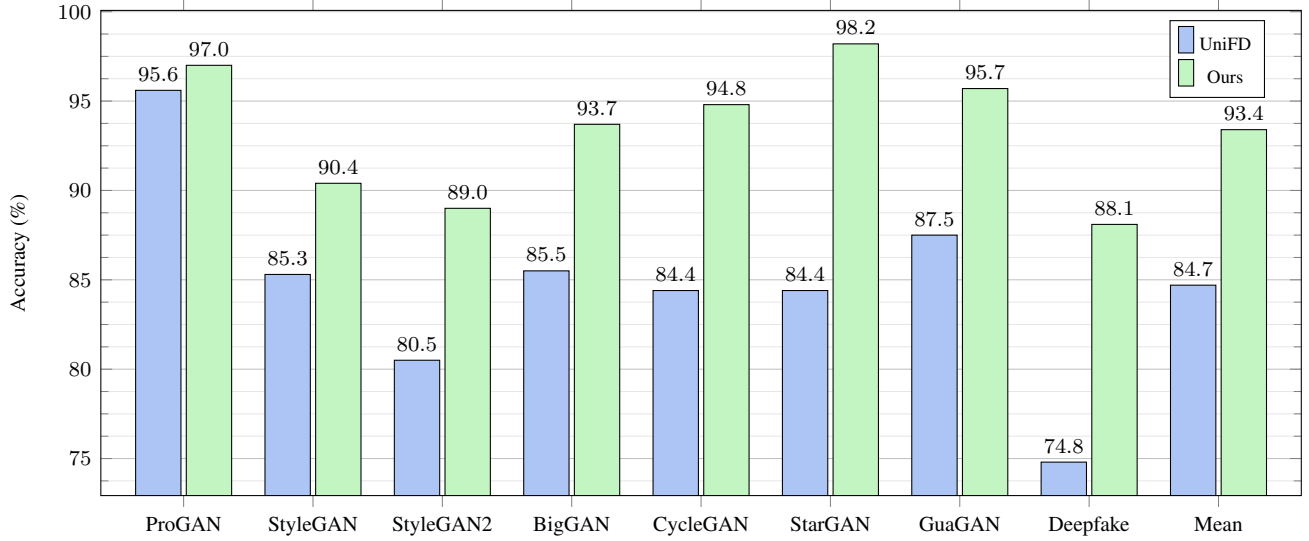
Figure 5. **Robustness comparisons with combined four image perturbations.** We report the accuracy results on the GANs dataset. By considering the forgery adaptation, our FatFormer works better on all generative models than UniFD which adopts the fixed pre-trained paradigm.