Supplementary Material for Forgery-aware Adaptive Transformer for Generalizable Synthetic Image Detection

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

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. Next, a comprehensive analysis is conducted to assess the efficacy of forgery adaptation in enhancing *robustness* (outlined in Section A.3) against image perturbations. Lastly, we provide the inference *efficiency* analysis in Section A.4.

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.

*Work done when H. Liu is a long-term intern at Baidu.
†Corresponding author (E-mail: yzhao@bjtu.edu.cn).

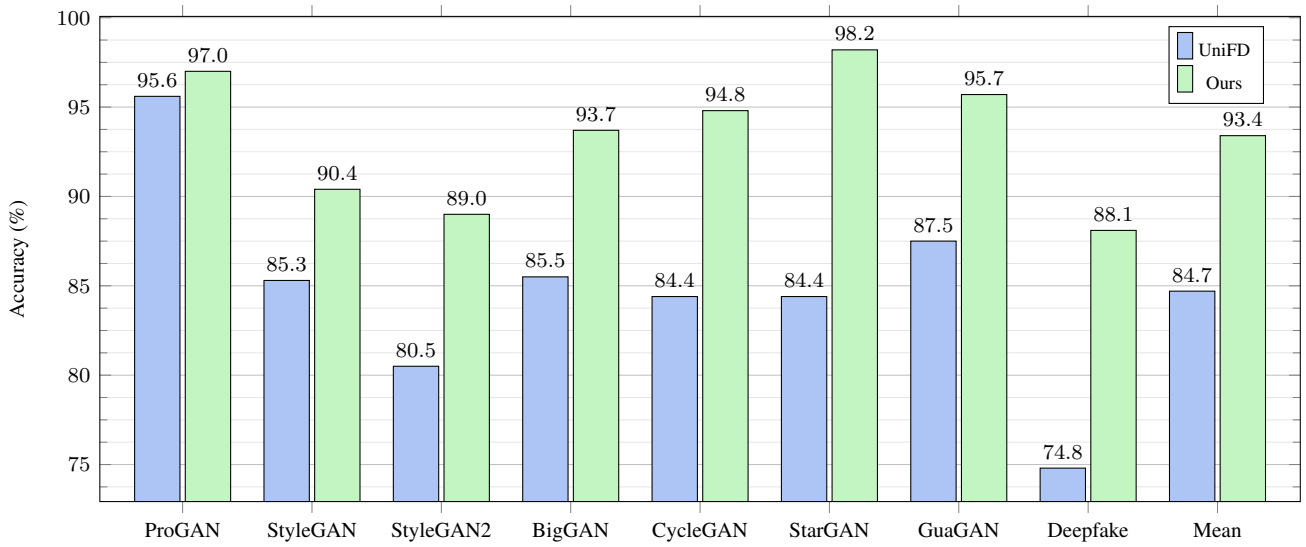

Figure 5. **Robustness comparisons with combined four image perturbations.** We report the accuracy results on the GANs dataset. By considering the forgery adaptation, our FatFormer works better on all generative models than UniFD which adopts the fixed pre-trained paradigm.

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, 55]. 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 [43], which adopts the fixed pre-trained paradigm. The results are shown in the following table:

Perturbation	Method	ACC_M	AP_M
Gaussian blurring	UniFD	78.1	93.0
	FatFormer	**90.7**	**98.1**
random cropping	UniFD	88.9	98.1
	FatFormer	**98.2**	**99.7**
JPEG compression	UniFD	88.4	97.7
	FatFormer	**95.9**	**99.2**
Gaussian noising	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.

A.4. Efficiency evaluation

The following table provides the comparison of inference FPS and inference time on the GANs dataset, based on NVIDIA A100 GPU.

Method	ACC_M	AP_M	FPS↑	Time [ms]↓
UniFD	89.1	98.3	104	9.6
FatFormer	95.3	99.5	110	9.1

We can see that FatFormer (based on CLIP ViT-B) achieves much better results and similar real-time inference speed than UniFD [43].