Forensic-MoE: Exploring Comprehensive Synthetic Image Detection Traces with Mixture of Experts

Mingqi Fang[1,2], Ziguang Li[1], Lingyun Yu[1,*], Quanwei Yang[1], Hongtao Xie[1], Yongdong Zhang[1,2]
[1]University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{mqfang, zg321, yangquanwei}@mail.ustc.edu.cn, {yuly, htxie, zhyd73}@ustc.edu.cn

In this supplementary material, we first report more experiment results in Section A. Afterwards, the details of the compared methods during experiment are provided in Section B. Moreover, the potential social impacts of our method are discussed in Section C. Finally, in Section D we discuss the method limitation and future work.

A. More Experiment Results

A.1. Expert Combinations

In Section 4.4 of the main text, we preliminarily discuss the expert selection problem about the number and type of experts. The law of ***"2-Experts, GAN+DM"*** is observed. To ensure comparison fairness, we directly adopt the widely used ProGAN and SDv1.4 experts during implementation. Table 1 further discusses the performance of diverse expert combinations with the mean Acc score. We find the combinations of ***"ProGAN + SDv1.4/SDv1.5"*** get satisfactory results. Moreover, other combinations also maintain a relatively stable detection performance.

A.2. Analysis with Different Architectures

In Table 2, we discuss the influence of different CLIP backbone architectures. We replace the default ViT-L/14 architecture with the ViT-B/16 architecture. Compared to directly training a universal detector, the utilization of our MoE paradigm also achieves consistent improvements with a smaller backbone, which proves the flexibility of our method.

Table 1. **Influence of Expert Combinations**

	SDv1.4	Midjourney	SDv1.5	Wukong
ProGAN	95.14	94.62	95.21	94.42
	ProGAN	StyleGAN	StyleGAN2	CycleGAN
SDv1.4	95.14	94.32	93.19	92.92

Table 2. **Analysis with different architecture.**

Arch	w/ MoE	Acc_M	AP_M
ViT-L/14	✗	93.78	98.89
	✓	**95.14**	**99.32**
ViT-B/16	✗	90.82	97.93
	✓	92.23	98.25

A.3. Influence on Each Expert after Finetuning

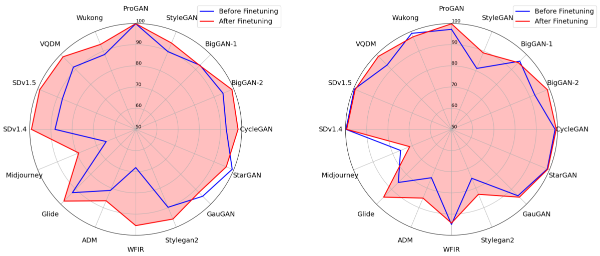
In Forensic-MoE, after the expert training step, we propose to enhance the expert interaction through the *MoE Finetuning* step and expert knowledge distillation. Each expert is encouraged to learn from others for two purposes, *i.e.,* (1) self-improvement of every expert and (2) enhancing the expert collaboration. In Table 3 of the main text, we have already discussed the benefit of our design in expert interaction and collaboration. In this section, we further discuss the influence on each expert's self-improvement.

Figure 1 compares each expert's performance before and after finetuning. Specifically, to evaluate the expert after finetuning, we freeze it and train an additional FC layer classifier with the corresponding images. It can be observed that, whether for the ProGAN or SDv1.4 experts, they both basically maintain the original discriminative within their own categories (GAN-family or diffusion-family), and they also uniformly achieve performance improvement in the other category after finetuning. It proves that the MoE finetuning step brings remarkable improvement for every individual expert.

A.4. Resource Overhead

Table 3 reports our resource overhead with ViT-L/14 and ViT-B/16 backbones. Different from the compared FatFormer, we only use the single image encoder rather than the whole CLIP model for synthetic detection. Moreover, benefiting from the flexible adapter-based structure, multiple experts will not introduce significant resource overhead, and it is suitable for real-world implementation.

[*] Corresponding author



(a) ProGAN Expert

(b) SDv1.4 Expert

Figure 1. **Comparison of expert performance before and after finetuning.**

Table 3. **Resource Overhead**

	Params (M)	FLOPs (G)	Inference Time(s)	Acc_M
FatFormer	492.59	109.57	0.2944	82.93
Ours(ViT-B/16)	212.38	47.32	0.0532	92.23
Ours(ViT-L/14)	520.50	190.36	0.1156	95.14

A.5. AP Score Comparison Result

In Table 4, we report the corresponding AP result of the experiment in the main text Section 4.2. Similar to the Acc score, it can be observed that our Forensic-MoE also achieves the best mean AP score across all the test sets.

B. Compared Methods

In Section 4.2 of the main text, we conduct generalization comparison with existing state-of-the-art methods, including CNNSpot [8], GramNet [3], LGrad [6], UnivFD [4], DIRE [9], FatFormer [2], and NPR [7]. Following, we provide the detailed introduction of each compared method.

It is worth noting that, some of the above methods use the UnivFD dataset [4] for Diffusion-family detection performance evaluation in their original papers. However, even with the same generative model, the GenImage dataset[10] is more challenging due to its broader subject coverage and larger data volume. Therefore, we select the GenImage dataset for evaluation in the main text.

- **CNNSpot** [8] proposes to detect synthetic images through deep learning models, and discusses the impact of image processing methods on detection results.
- **GramNet** [3] suggests focusing on the global texture of the image for effective detection.
- **LGrad** [6] finds that the gradient information in generative models like StyleGAN [1] reveals the forensic information and can be utilized for detection.
- **UnivFD** [4] finds that benefit from the abundant prior knowledge of real image distribution, several large-scale pretrained models like CLIP [5] exhibit significant potential in synthetic detection. It only utilizes the simple nearest neighbor classification and achieves remarkable performance.
- **DIRE** [9] proposes that diffusion models generally can reconstruct better synthetic images than real images. Accordingly, the reconstruction error is utilized to distinguish the synthetic images.
- **FatFormer** [2] designs a forgery-aware adapter and text-guided contrastive learning schema to aggregate forensic traces, thus further improving the CLIP-based synthetic detection performance.
- **NPR** [7] suggests focusing on the up-sampling generation artifact and designing a neighboring pixel relationships feature for generalizable synthetic detection.

C. Potential Social Impacts

Although developing synthetic image detection tools is beneficial for preventing the spread of misinformation and guaranteeing visual content trustworthiness. However, it may still lead to misjudgments and hinder the dissemination of normal content. Firstly, a detection system deployed on the website server may mistakenly identify and intercept the authentic content, resulting in unnecessary information spread costs. Moreover, the detection results could be utilized as evidence in legal proceedings, and the misclassified cases may mislead the legal judgment. Before developing the detection tools for content moderation and evidence collection, these potential social impacts should be thoroughly considered.

D. Limitations and Future Works

Our mixture of experts architecture exhibits remarkable detection performance on diverse synthetic methods. In real-world deployment, the addition of every new expert requires the whole model to be finetuned again, which can be improved to enhance flexibility. The investigation of this problem is left in future work. We hope to simplify the finetuning process and enhance the real-world deployment flexibility through the incremental learning paradigm.

References

[1] 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

[2] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10770–10780, 2024. 2, 3

[3] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Pro-*

Table 4. **AP score comparison.** We report the AP score of each method. BigGAN[1] and BigGAN[2] respectively refer to the subset from ForenSynths and GenImage datasets. † represents that the model is trained with only GAN images according to the original papers. Red and Blue represent the best and second-best performance respectively. The mean AP is marked in gray .

Test Sets (AP%)		CNNSpot [8]	GramNet [3]	LGrad [6]	UnivFD [4]	DIRE [9]	FatFormer† [2]	FatFormer [2]	NPR† [7]	NPR [7]	Ours
GAN-family	ProGAN	100.00	100.00	100.00	99.90	100.00	100.00	100.00	99.99	100.00	100.00
	StyleGAN	98.01	95.99	98.89	93.26	99.46	99.75	99.91	99.78	98.29	99.40
	StyleGAN2	97.71	96.04	98.97	92.98	99.98	99.92	99.56	99.94	98.69	99.59
	BigGAN[1]	71.82	66.60	77.32	94.69	64.00	99.98	98.94	87.80	73.98	97.92
	BigGAN[2]	93.76	85.88	95.07	99.62	99.97	99.54	99.84	88.08	99.97	99.86
	CycleGAN	91.49	79.03	97.37	99.06	70.08	100.00	99.94	98.45	97.27	99.83
	StarGAN	97.72	99.98	99.84	98.77	99.93	100.00	100.00	99.94	80.36	99.86
	GauGAN	91.59	52.86	69.86	99.00	60.84	100.00	99.85	85.49	77.45	98.29
	WFIR	52.64	55.00	51.89	93.71	52.86	98.48	79.23	65.32	66.26	99.85
	Mean	88.30	81.26	87.69	96.78	83.01	99.74	97.47	91.64	88.03	99.40
Diffusion-family	Midjourney	89.92	86.49	93.61	95.76	99.91	62.76	95.88	85.36	99.42	96.54
	SDv1.4	99.99	99.97	99.96	97.99	100.00	81.12	100.00	84.03	100.00	99.93
	SDv1.5	99.98	99.95	99.92	97.70	100.00	81.09	99.97	84.59	99.93	99.87
	ADM	70.84	71.07	70.30	85.39	99.99	91.73	83.42	74.64	75.40	98.72
	GLIDE	85.11	83.04	95.71	97.52	99.99	95.99	99.68	85.72	97.79	99.72
	Wukong	99.99	99.79	99.78	96.49	100.00	85.86	100.00	80.47	99.96	99.90
	VQDM	56.40	60.66	76.99	93.14	100.00	97.26	96.13	82.89	78.16	99.79
	Mean	86.03	85.85	90.90	94.86	99.98	85.12	96.44	82.53	92.95	99.21
Overall Mean		87.31	83.27	89.09	95.94	90.44	93.34	97.02	87.66	90.18	99.32

ceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8060–8069, 2020. 2, 3

[4] 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. 2, 3

[5] 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. 2

[6] 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. 2, 3

[7] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024. 2, 3

[8] 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. 2, 3

[9] 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*, pages 22445–22455, 2023. 2, 3

[10] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36, 2024. 2