Forging the Unknown: Open-Set Deepfake Attribution via Adaptive Fingerprint Learning

Yizhi Fang[1][0009-0001-4905-1795], Boxuan Han[1][0009-0004-2528-1141], Jingwen Wang[1][0009-0007-3719-3116], Xiandang Luo[2][0009-0004-8294-6875], Siyu Peng[1][0009-0004-2528-1141], Xiarun Chen[1][0009-0002-1789-4584], Weiping Wen*[1][0009-0002-6313-2943], and Sai Cheng[3][0000-0002-5799-5622]

[1] Peking University, Beijing 100871, China
[2] Xiangtan University, Xiangtan 411100, China
[3] Xiaomi Corporation, Beijing 100871, China

Abstract. With the rapid advancement of Generative AI, deepfake generators are iterating and proliferating at an unprecedented pace, making "zero-day" forgery attacks a routine challenge in digital forensics. However, existing deepfake attribution methods largely rely on the closed-set assumption, i.e., test samples are generated only by the generators seen during training. Under this setting, models often exhibit severe over-confidence when confronted with samples from unseen generators, wrongly assigning them to a known class with high confidence. To overcome this limitation, we propose a novel Open-Set Deepfake Attribution (OSDA) framework. The core of our approach is an Adaptive Fingerprint Learning (AFL) strategy, aiming to learn more discriminative and content-agnostic generator "fingerprints". Specifically, we design a spatial–frequency dual-stream network to jointly extract subtle artifacts from spatial residuals and the spectral domain. We further introduce Adaptive Margin Metric Learning to enforce intra-class compactness and inter-class separability on a hyperspherical feature manifold, explicitly reserving open space for unknown classes. During inference, we adopt an Extreme Value Theory (EVT)-based distance modeling mechanism to enable statistically reliable rejection of unknown samples. Experiments on GenImage and FaceForensics++ demonstrate that OSDA significantly outperforms prior methods in both closed-set attribution accuracy and open-set AUROC, and exhibits excellent robustness to post-processing operations such as JPEG compression and Gaussian blur.

Keywords: Deepfake attribution · Open-set recognition · Generator fingerprint · Metric learning

* Corresponding author.

1 Introduction

1.1 Background

The emergence of Generative Artificial Intelligence (GenAI) has fundamentally transformed digital content creation. Advancements in deep generative modeling, evolving from early Generative Adversarial Networks (GANs) [1] to contemporary Latent Diffusion Models (LDMs) [2], have pushed the visual quality of synthetic media to a level often indistinguishable to the human eye. Consequently, high-fidelity forged content can now be produced at scale, lowering the technical barrier for creative expression and malicious misuse alike. As these technologies become ubiquitous, the field of digital forensics faces an increasingly complex landscape of synthetic artifacts [3].

1.2 Motivation

Despite the creative potential of GenAI, the proliferation of deepfakes poses severe risks to information integrity, including the spread of misinformation and the infringement of individual rights. Traditional detection frameworks are often designed for closed-set scenarios, where the detector is trained on specific, known forgery algorithms. However, these methods frequently fail when encountering novel generative distributions or "in-the-wild" content.

A critical turning point occurred in 2023, when a forged video of a political figure—generated by a then-undisclosed diffusion model—successfully bypassed mainstream detection systems. This incident underscores the fragility of existing countermeasures against unseen generator architectures. The motivation for this work is to bridge this security gap by developing a more robust detection mechanism capable of identifying diverse synthetic signatures, particularly those originating from advanced diffusion processes that deviate from GAN-based patterns.

1.3 Problem Statement

To address these threats, the multimedia forensics community has developed various detection methods. While binary detection (real vs. fake) constitutes the first line of defense [4–7], deepfake attribution—identifying the specific generation source (e.g., ProGAN, StyleGAN, Stable Diffusion)—is equally crucial for accountability and judicial investigations [8, 9].

However, most existing attribution techniques rely on the closed-set assumption [10], where the training and test sets share exactly the same label space. This assumption has a fundamental flaw in practice: in real-world environments, the ecosystem of generative models evolves dynamically, with new architectures and "zero-day" forgeries continuously emerging. When a conventional closed-set classifier encounters a sample from an unknown generator, it tends to force it into one of the known classes, often with high confidence. This over-confidence phenomenon severely undermines the reliability of forensic systems [11].

1.4 Challenges

Addressing this issue requires shifting the paradigm from closed-set classification to Open-Set Deepfake Attribution (OSDA) [10]. OSDA has a dual objective: accurately classify samples from known generators, and effectively reject samples from unknown sources [11, 12]. Compared with general Open-Set Recognition (OSR) tasks, OSDA introduces unique challenges: deepfake "fingerprints" are subtle high-frequency artifacts that are easily affected by semantic content or destroyed by post-processing operations (e.g., compression) [4, 6]. Unlike natural image classification, where classes such as "cats" and "dogs" differ semantically, fake images generated by different GANs or diffusion models are often visually similar in content. Their differences mainly lie in sub-pixel, high-frequency noise patterns [8, 13], imposing higher demands on feature extractor design. Generic OSR methods typically rely on semantic features and therefore struggle to capture such fine-grained, content-agnostic traces, resulting in poor separation between manifolds of known and unknown generators [14].

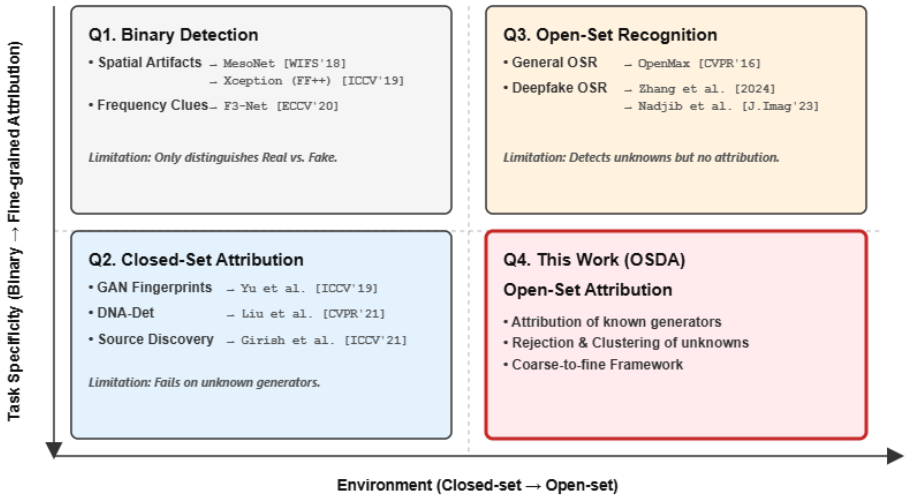


Fig. 1: The evolution of deepfake attribution tasks: from binary detection (Q1), closed-set attribution (Q2), open-set recognition (Q3), to the open-set attribution (Q4) proposed in this work.

1.5 Solution

To overcome these challenges, we propose a novel OSDA framework driven by Adaptive Fingerprint Learning (AFL). Our key insight is that reliable open-set detection requires a feature space that explicitly disentangles generator "fingerprints" from image content [8, 9]. To this end, we design a dual-stream network

that works collaboratively in complementary domains: the spatial stream captures pixel-level anomalies via residual learning [4], while the frequency stream reveals artificial traces introduced by up-sampling or diffusion processes through spectral analysis [6, 15–17]. By fusing features from both streams, we obtain a comprehensive representation of intrinsic generator signatures.

1.6 Technical Innovations

However, feature extraction alone is insufficient to handle unknown classes; the embedding space must also be structurally designed. We introduce an Adaptive Margin Metric Learning objective inspired by large-margin cosine/angular metric learning [18–20]. Unlike the standard Softmax loss, this objective dynamically adjusts decision boundaries to ensure high intra-class compactness and maximal inter-class separation. Crucially, it explicitly reserves "open space" in the feature manifold for potential unknown classes, consistent with open-set risk control [10]. During inference, we abandon naive thresholding and instead incorporate Extreme Value Theory (EVT) [11, 21] to model the tail distribution of distances to known class centroids, providing a statistically principled rejection mechanism for zero-day attacks.

1.7 Main Contributions

In summary, our main contributions are:

- We expose a critical over-confidence issue in existing attribution models and propose a complete Open-Set Deepfake Attribution (OSDA) framework to address zero-day forgery attacks [10, 11];
- We design an Adaptive Fingerprint Learning (AFL) strategy, including a spatial–frequency dual-stream fusion architecture [4, 6, 16] and an adaptive margin metric learning objective [18, 19], to build a robust and separable embedding space;
- Extensive experiments on GenImage [22] and FaceForensics++ [3] demonstrate that our method achieves state-of-the-art performance in both closed-set attribution accuracy and open-set AUROC, and shows strong robustness to common perturbations (e.g., JPEG compression and blur).

2 Related Work

This section reviews three research areas most relevant to this paper: deepfake attribution, open-set recognition, and their intersection.

2.1 Deepfake Detection and Attribution

Early studies mainly focused on binary detection, aiming to distinguish real from fake media. Methods include leveraging spatial artifacts [4, 13, 5, 23–25], frequency inconsistencies [6, 16, 17, 15], and biological signals [7].

With the rapid growth of generative models, merely detecting authenticity is no longer sufficient for forensic needs, and attribution has emerged accordingly. Yu et al. first proposed fingerprint-based attribution methods that identify GAN sources using multi-class CNNs [8]. Subsequent studies further improved attribution accuracy in closed-set settings by leveraging global texture statistics [26] and DNA-Det [27]. Related works also explore training-data-rooted fingerprints and source identification settings for generative media [9, 28].

Although the above methods perform well under closed-set settings, they all assume that generators in the test set are known during training. When faced with unknown generators (open-set), these models tend to force unknown samples into the most similar known class with dangerously high over-confidence [10, 11].

2.2 Open-Set Recognition

Open-set recognition aims to train models that correctly classify known classes while rejecting unknown classes [10]. Classic OSR methods include OpenMax [11], which fits a Weibull distribution to calibrate Softmax scores, and prototype- or distance-based approaches that constrain the geometry of feature space. Recent competitive OSR methods include ARPL [12], as well as approaches emphasizing strong closed-set classifiers for OSR [14] and class-conditional density modeling [29]. Generative variants such as Generative OpenMax have also been explored [30].

Generic OSR methods usually rely on salient semantic features (e.g., distinguishing "cats" from "dogs"). However, deepfake differences mainly reside in subtle, content-agnostic high-frequency fingerprints [8, 4]. Directly transferring generic OSR methods often fails for deepfake tasks because the feature extractor focuses on image content rather than generation traces.

2.3 Open-Set Forensics

Only a limited number of works have explored open-set multimedia forensics. Some pioneering studies attempt to use Bayesian uncertainty or ensemble learning, but these approaches are often computationally expensive and struggle to maintain efficient feature separation for fine-grained generator classification. Recently, Zhang et al. [31] introduced open-set ideas into deepfake detection, but focused on binary real/fake discrimination rather than fine-grained generator attribution. This work is the first to systematically define and address the open-set deepfake attribution problem, filling this gap.

3 Methodology

3.1 Problem Formulation

Let the training dataset be $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is an input image and $y_i \in \mathcal{K} = \{1, \ldots, K\}$ is the label of one of the K known generators.

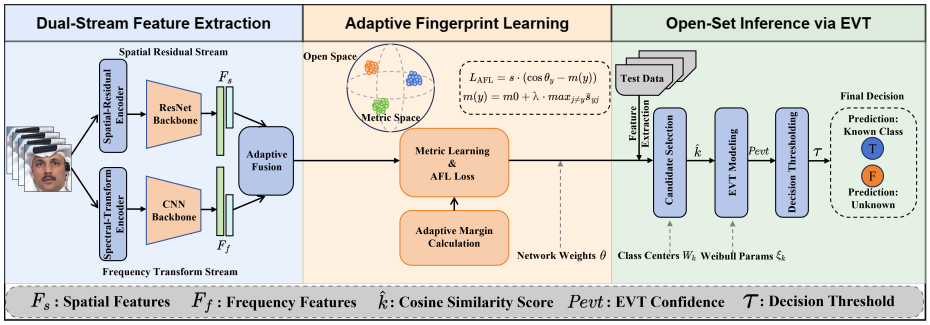


Fig. 2: Overall framework of the proposed method: we first extract generator fingerprint features F_s and F_f from the spatial residual stream and the frequency spectral stream, respectively, and adaptively fuse them via attention gating to obtain the embedding representation; then, in the Adaptive Fingerprint Learning (AFL) stage, we employ an adaptive margin metric learning loss to compress intra-class distributions and enlarge inter-class margins, reserving open space for unknown generators; during inference, we select candidate classes based on cosine similarity to class centers and use Extreme Value Theory (EVT) to fit a Weibull model to the tail distribution, compute P_{evt}, and compare it with a threshold τ to achieve known-class recognition and unknown-class rejection.

At test time, samples come from $\mathcal{D}_{\text{test}}$, whose label space expands to $\mathcal{Y}_{\text{test}} = \mathcal{K} \cup \mathcal{U}$, where $\mathcal{U}$ denotes unknown generator classes, following the standard open-set formulation [10].

Our goal is to learn a mapping function $f(x; \theta)$ such that:

(1) If a sample belongs to a known class $k \in \mathcal{K}$, it is correctly classified: $\hat{y} = k$.
(2) If a sample belongs to an unknown class $u \in \mathcal{U}$, it triggers rejection: $\hat{y} =$ unknown.

3.2 Dual-Stream Feature Extraction

To extract robust, content-agnostic generation fingerprints, we design a dual-stream network motivated by prior evidence that both spatial artifacts and frequency-domain cues are effective for deepfake detection and attribution [4, 6, 16, 17].

Spatial Residual Stream: Generator traces are often hidden in the high-frequency residuals of images [4, 13]. Instead of using raw RGB images directly, we preprocess the input through a learnable residual module:

$$x_{\text{res}} = x - \text{Smooth}(x) \tag{1}$$

where Smooth$(\cdot)$ can be implemented as a 5×5 mean filter, or a learnable SRM (Spatial Rich Model) filter bank [32], which more effectively captures high-

frequency anomalies left by common image processing. This stream extracts spatial features $F_s \in \mathbb{R}^d$ using a ResNet backbone [33].

Frequency Spectral Stream: Up-sampling operations often leave distinctive checkerboard artifacts in the frequency domain [15]. We transform the image into the frequency domain (using DCT or FFT) and use its magnitude spectrum as input. This stream extracts frequency features $F_f \in \mathbb{R}^d$ using a CNN, following frequency-aware deepfake analysis [6, 16, 17].

Adaptive Fusion: To combine the complementary strengths of both streams, we employ an attention mechanism for feature fusion:

$$F_{\text{fused}} = \alpha \cdot F_s + (1 - \alpha) \cdot F_f \tag{2}$$

where α is a dynamically learned weighting parameter produced by a lightweight gating network.

3.3 Adaptive Fingerprint Learning

This is the core of our method. The standard Softmax loss focuses only on classification accuracy and does not enforce intra-class compactness, resulting in inefficient use of feature space and causing unknown samples to fall into decision regions of known classes, which is problematic in OSR [10, 11].

We introduce Adaptive Margin Metric Learning, inspired by large-margin embedding learning methods such as ArcFace [18], CosFace [19], and AdaCos [20].

We adopt a cosine-distance-based loss with an adaptive margin m. Let the feature vector z_i and the class center W_{y_i} be normalized. The loss is defined as:

$$L_{\text{AFL}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{s(\cos \theta_{y_i} - m(y_i))}}{e^{s(\cos \theta_{y_i} - m(y_i))} + \sum_{j \neq y_i} e^{s \cos \theta_j}} \tag{3}$$

where:

- s is the scale factor;
- θ_{y_i} is the angle between feature z_i and its ground-truth class center W_{y_i};
- $m(y_i)$ is the adaptive margin. Unlike the fixed m in ArcFace [18], we dynamically adjust m based on class difficulty. Specifically, we estimate the confusion level based on the average cosine similarity between classes during training:

$$\bar{s}_{ij} = \frac{1}{N_i N_j} \sum_{x_a \in C_i, x_b \in C_j} \cos(z_a, z_b). \tag{4}$$

  For class y_i, the adaptive margin is defined as:

$$m(y_i) = m_0 + \lambda \cdot \max_{j \neq i} \bar{s}_{ij}, \tag{5}$$

  where $m_0 = 0.5$ is the baseline margin and $\lambda = 0.3$ is the adjustment coefficient. Higher similarity implies a larger margin, enforcing stronger separation.

In this way, known classes are compressed into compact regions on the hypersphere, reserving substantial open space in the manifold for unknown classes, consistent with open-set risk considerations [10].

3.4 Open-Set Inference via EVT

During inference, we do not rely on Softmax probabilities (which suffer from normalization-induced over-confidence), but instead make decisions based on **distance** and EVT calibration [11, 21].

1. **Feature extraction:** Compute the embedding feature z^* for a test sample x^*.
2. **Distance computation:** Compute the cosine distance between z^* and all known class centers W_k. Select the nearest class as the candidate prediction:
$$\hat{k} = \arg\max_{k \in \mathcal{K}} \cos(z^*, W_k) \tag{6}$$
3. **Extreme value modeling:** To decide whether to reject, we leverage EVT. After training, we collect the "tail distance" distribution for each known class from correctly classified samples (i.e., the top 20 samples with the largest distances to the class center) and fit a Weibull distribution using maximum likelihood estimation (MLE), following OpenMax-style calibration [11]. Given the distance between the test sample and class $\hat{k}$, we compute its probability under this distribution, denoted as P_{evt}.
4. **Final decision:** Given a threshold τ:
$$\text{Prediction} = \begin{cases} \hat{k}, & \text{if } P_{\text{evt}} > \tau \\ \text{Unknown}, & \text{otherwise} \end{cases} \tag{7}$$

4 Experiments

4.1 Experimental Setup

Datasets & Protocols: We evaluate on GenImage [22] and FaceForensics++ (FF++) [3]. For GenImage, we adopt an open-set protocol with a random 4-known/4-unknown split and average over multiple repeats. For FF++, we use a 3-known/1-unknown leave-one-out protocol and report the mean and standard deviation across 4 runs.

Evaluation Metrics: We report closed-set attribution accuracy on known classes and open-set metrics including AUROC, FPR@95TPR, and OSCR [10].

Implementation Details: We use ResNet-50 (ImageNet pretrained) [33]. Images are resized to 256×256 and randomly cropped to 224×224 during training. We train with SGD (initial lr 0.01, momentum 0.9, weight decay 10^{-4}) for 50 epochs (batch size 64), decaying lr by 0.1 at epochs 30 and 40. Hyperparameters: s=30, m_0=0.5, λ=0.3, EVT tail size T=20.

4.2 Main Results

We compare our method with MSP (maximum softmax probability), OpenMax [11], and ARPL [12]. The results on GenImage are summarized in Table 1, and the leave-one-out evaluation on FaceForensics++ is reported in Table 2.

Table 1: Open-set attribution results on GenImage (4 known / 4 unknown).

Method	Closed-set Acc (%) ↑	AUROC (%) ↑	FPR@95TPR (%) ↓	OSCR (%) ↑
MSP (Softmax)	98.2	65.4	61.8	60.7
OpenMax	97.5	72.1	49.6	67.9
ARPL	98.5	81.3	31.4	77.2
Ours (OSDA)	**99.1**	**92.6**	**12.7**	**89.8**

Table 2: Leave-one-out open-set results on FaceForensics++ (AUROC %).

Method	DF	F2F	FS	NT	Average
MSP (Softmax)	78.2	75.6	72.1	76.8	75.7 ± 2.6
OpenMax	82.3	80.1	78.5	81.2	80.5 ± 1.6
ARPL	86.7	85.2	84.0	85.9	85.5 ± 1.1
Ours (OSDA)	**93.1**	**91.8**	**90.5**	**92.3**	**91.9 ± 1.0**

Across both benchmarks, OSDA achieves strong closed-set attribution and substantially improved open-set discrimination, validating the benefit of adaptive metric learning and EVT-based rejection.

4.3 Stronger Baselines

We further compare with common open-set detectors built on the same ResNet-50 features, including Energy score, ODIN, Mahalanobis distance, and kNN-OSR, following the same data splits and known-only validation for threshold selection. As shown in Table 3, our method consistently outperforms these strong baselines across all metrics.

4.4 Ablation Studies

We conduct ablation studies on GenImage to verify the effectiveness of each component. Table 4 shows the contribution of the spatial stream, frequency stream, metric loss, and EVT-based rejection. Furthermore, we analyze different designs of the metric learning module in Table 5.

Table 3: Extended comparison on GenImage (4 known / 4 unknown).

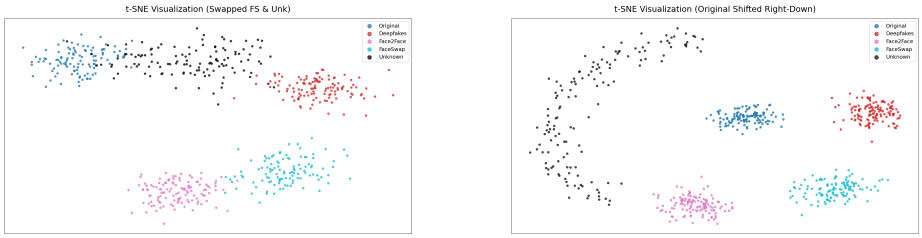
Method	Closed-set Acc (%) ↑	AUROC (%) ↑	FPR@95TPR (%) ↓	OSCR (%) ↑
MSP (Softmax)	98.2	65.4	61.8	60.7
Energy score	98.0	68.9	57.3	64.1
ODIN	98.1	70.2	54.8	65.5
Mahalanobis	97.8	73.5	48.2	69.0
kNN-OSR	97.9	75.6	45.1	71.3
OpenMax [11]	97.5	72.1	49.6	67.9
ARPL [12]	98.5	81.3	31.4	77.2
Ours (OSDA)	**99.1**	**92.6**	**12.7**	**89.8**

Table 4: Ablation study of key components (GenImage, AUROC %).

Variant	Spatial	Freq.	Metric Loss	EVT	AUROC (%) ↑
Baseline	✓				68.5
+ Freq	✓	✓			75.2
+ Metric	✓	✓	✓		88.4
Full Model	✓	✓	✓	✓	**92.6**

The frequency stream improves open-set separation, metric learning yields the largest gain, and EVT further boosts stability by calibrating tail risk for rejection decisions.

4.5 Visualization



(a) MSP (Softmax) Baseline (b) Ours (OSDA)

Fig. 3: t-SNE feature visualization comparison: (a) MSP baseline; (b) the proposed OSDA method.

Figure 3 shows that OSDA learns more compact and well-separated clusters for known generators, while pushing unknown samples into open regions, consistent with the improvements in open-set metrics.

Table 5: Fine-grained ablation on metric learning design (GenImage, AUROC %).

Variant	AUROC (%) ↑
Fixed margin (ArcFace-style)	89.7
Adaptive margin (Ours)	**92.6**
Adaptive margin w/o dynamic adjustment (λ=0)	90.8
Adaptive margin w/o margin initialization (remove m_0)	91.1

5 Conclusion

5.1 Summary

This paper addresses the "zero-day attack" challenge in deepfake attribution and proposes a new Open-Set Deepfake Attribution (OSDA) framework [10, 11]. We identify the limitations of traditional closed-set methods and introduce an Adaptive Fingerprint Learning (AFL) strategy. By combining a spatial–frequency dual-stream network [4, 6] with adaptive margin metric learning [18, 19], our model not only captures content-agnostic generation fingerprints accurately, but also reserves sufficient embedding space for unknown classes. Together with an EVT-based inference mechanism [11, 21], the proposed approach enables accurate attribution of known generators and effective rejection of unknown generators.

5.2 Limitations

Although OSDA exhibits superior performance in detecting unknown generators, we acknowledge the following limitations, which also point to future research directions:

1. **Confusion of fine-grained homologous variants:** When an unknown generator shares highly similar architecture or pretrained weights with a known generator (e.g., Stable Diffusion v1.4 vs. v1.5, or different fine-tuned checkpoints of the same GAN), their frequency fingerprints can be extremely similar [2]. In such high-homology scenarios, adaptive metric learning may fail to fully separate them in feature space, causing unknown samples to be misclassified as homologous variants of known classes.
2. **Vulnerability to adversarial washing:** Our method primarily relies on high-frequency artifacts left by the generation process [4, 6]. If an attacker uses adversarial example techniques or specialized denoising post-processing (anti-forensic noise removal) to deliberately erase these fingerprints, the feature extraction capability of the dual-stream network may be suppressed, potentially reducing sensitivity to carefully camouflaged deepfakes.
3. **Failure under severe degradation:** While experiments show robustness to standard JPEG compression, under extreme degradation (e.g., low-bitrate

video after repeated transcoding on social platforms, screen re-capture, or low-resolution thumbnails), high-frequency information may be almost entirely lost. In such cases, the discriminative power of the spectral stream may significantly degrade, and the model may collapse to relying mainly on spatial semantics, increasing the miss rate for unknown attacks.

5.3 Future Work

Future work will focus on:

1. **Fine-grained unknown clustering:** Beyond detecting "unknown", cluster unknown samples to discover newly emerging generator families.
2. **Incremental learning:** Study how to register newly discovered unknown generators as known classes quickly without retraining the entire model.
3. **Cross-modal generalization:** Explore whether an OSDA model trained on images can be directly applied to video frames for open-set attribution, supporting real-time forged-content monitoring systems.

References

1. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
2. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
3. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Niessner, M.: FaceForensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1–11 (2019)
4. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: CNN-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8695–8704 (2020)
5. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: MesoNet: A compact facial video forgery detection network. In: IEEE International Workshop on Information Forensics and Security, pp. 1–7 (2018)
6. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: Proceedings of the International Conference on Machine Learning, pp. 3247–3257 (2020)
7. Ciftci, U.A., Demir, I., Yin, L.: FakeCatcher: Detection of synthetic portrait videos using biological signals. IEEE Transactions on Pattern Analysis and Machine Intelligence, **44**(5), 2368–2379 (2022)
8. Yu, N., Davis, L.S., Fritz, M.: Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7556–7566 (2019)
9. Yu, N., Davis, L.S., Fritz, M.: Rooting deepfake attribution in training data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3725–3734 (2021)

10. Scheirer, W.J., de Rezende Rocha, A., Sapkota, A., Boult, T.E.: Toward open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, **35**(7), 1757–1772 (2013)
11. Bendale, A., Boult, T.E.: Towards open set deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1563–1572 (2016)
12. Chen, G., Peng, P., Wang, X., Tian, Y.: Adversarial reciprocal points learning for open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, **44**(11), 8065–8081 (2022)
13. Chai, L., Bau, D., Lim, S.N., Isola, P.: What makes fake images detectable? Understanding properties that generalize. In: Proceedings of the European Conference on Computer Vision, pp. 103–120 (2020)
14. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Open-set recognition: A good closed-set classifier is all you need. In: Proceedings of the International Conference on Learning Representations (2022)
15. Durall, R., Keuper, M., Keuper, J.: 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, pp. 7887–7896 (2020)
16. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: Proceedings of the European Conference on Computer Vision, pp. 86–103 (2020)
17. Luo, Y., Zhang, Y., Yan, J., Liu, W.: Generalizing face forgery detection with high-frequency features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16317–16326 (2021)
18. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)
19. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Liu, T.: CosFace: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5265–5274 (2018)
20. Zhang, X., Zhao, R., Qiao, Y., Wang, X., Li, H.: AdaCos: Adaptively scaling cosine logits for effectively learning deep face representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2714–2722 (2019)
21. Vignotto, E., Engelke, S.: Extreme value theory for open set classification—GPD and GEV classifiers. Statistics and Computing, **30**(6), 1485–1496 (2020)
22. Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Wang, Y.: GenImage: A million-scale benchmark for detecting and attributing generative images. In: Advances in Neural Information Processing Systems, **36**, 71957–71974 (2023)
23. Guarnera, L., Giudice, O., Battiato, S.: Deepfake detection by analyzing convolutional traces. IEEE Access, **8**, 130285–130295 (2020)
24. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., Yu, N.: Multi-attentional deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2185–2194 (2021)
25. Huang, Y., Juefei-Xu, F., Wang, R., Guo, Q., Ma, L., Xie, X., Liu, Y.: FakePolisher: Making deepfakes more detection-evading by removing local artifacts. In: Proceedings of the ACM International Conference on Multimedia, pp. 97–105 (2020)

26. Liu, Z., Qi, X., Torr, P.H.: Global texture enhancement for fake face detection in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8057–8066 (2020)
27. Yang, T., Li, X., Qi, H., Lyu, S.: DNA-Det: Detecting digital nature anomalies for deepfake video detection. IEEE Transactions on Information Forensics and Security, **17**, 1124–1138 (2022)
28. Kim, M., Tariq, S., Woo, S.S.: Decentralized source identification of deepfake videos. In: Proceedings of the AAAI Conference on Artificial Intelligence, **35**(15), 13409–13417 (2021)
29. Sun, X., Yang, Z., Zhang, C., Ling, K.V., Peng, G.: Conditional Gaussian distribution learning for open set recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13477–13486 (2020)
30. Ge, Z., Demyanov, S., Chen, Z., Garnavi, R.: Generative openmax for multi-class open set classification. In: Proceedings of the British Machine Vision Conference (2017)
31. Zhang, L., Yan, Q., Liu, Y.: Open-set deepfake detection via uncertainty modeling. In: Proceedings of the AAAI Conference on Artificial Intelligence, **38**(7), 6994–7002 (2024)
32. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital images. IEEE Transactions on Information Forensics and Security, **7**(3), 868–882 (2012)
33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)