AI-Generated Image Detection: An Empirical Study and Future Research Directions

Nusrat Tasnim[1]
tasnim.nishu70@kau.kr
Kutub Uddin[2]
kutub@umich.edu
Khalid Mahmood Malik[2]
drmalik@umich.edu

[1] School of Electronics and Information Engineering
Korea Aerospace University
Goyang, South Korea

[2] College of Innovation & Technology
University of Michigan-Flint,
Michigan, USA

Abstract

The threats posed by AI-generated media, particularly deepfakes, are now raising significant challenges for multimedia forensics, misinformation detection, and biometric system resulting in erosion of public trust in the legal system, significant increase in frauds, and social engineering attacks. Although several forensic methods have been proposed, they suffer from three critical gaps: (i) use of non-standardized benchmarks with GAN- or diffusion-generated images, (ii) inconsistent training protocols (e.g., scratch, frozen, fine-tuning), and (iii) limited evaluation metrics that fail to capture generalization and explainability. These limitations hinder fair comparison, obscure true robustness, and restrict deployment in security-critical applications. This paper introduces a unified benchmarking framework for systematic evaluation of forensic methods under controlled and reproducible conditions. We benchmark ten SoTA forensic methods (scratch, frozen, and fine-tuned) and seven publicly available datasets (GAN and diffusion) to perform extensive and systematic evaluations. We evaluate performance using multiple metrics, including accuracy, average precision, ROC-AUC, error rate, and class-wise sensitivity. We also further analyze model interpretability using confidence curves and Grad-CAM heatmaps. Our evaluations demonstrate substantial variability in generalization, with certain methods exhibiting strong in-distribution performance but degraded cross-model transferability. This study aims to guide the research community toward a deeper understanding of the strengths and limitations of current forensic approaches, and to inspire the development of more robust, generalizable, and explainable solutions.

1 Introduction

In recent times, the proliferation of AI-generated content, particularly deepfakes [57], has overwhelmed social media [1] and news platforms [7]. These deepfake contents are often used to mislead audiences by fabricating events or impersonating individuals, thereby undermining public trust. Moreover, deepfakes have emerged as critical threats to society, especially in security-sensitive domains. For instance, they can compromise biometrics used for face recognition and identification [55, 63], surveillance systems [26], and mislead perception modules in autonomous driving [18, 53]. Additionally, deepfakes pose significant risks in the Internet of Things (IoT) ecosystem [9, 54] and remote authentication systems [52, 62], where identity integrity is crucial. As the quality and realism of AI-generated content continue to improve, the ability to detect deepfake media has become increasingly challenging and making it imperative to develop robust and generalizable techniques.

Current statistics highlight the urgent need to verify deepfake content in domains such as social media, journalism, finance, the legal system, and governance. Across industries, deepfake fraud has become alarmingly common. Nearly 92% of companies have reported financial losses due to deepfake scams [8]. On average, businesses lose approximately $450,000 per incident, with the financial sector bearing even heavier losses, averaging $600,000 per organization, and in some cases exceeding $1 million [19]. In one notable incident, a Hong Kong employee was tricked into transferring $25 million after fraudsters used a deepfake video call to impersonate the company's CFO [14].

Globally, deepfake-enabled fraud caused over $200 million in losses during the first quarter of 2025 alone, indicating a rapidly escalating threat [16]. The cryptocurrency sector saw an even more dramatic impact, with deepfake-related scams increasing by 456% between May 2024 and April 2025, culminating in more than $10.7 billion in damages in 2024 [43]. Market projections suggest that generative AI-related fraud losses could rise to $40 billion by 2027, up from $12.3 billion in 2023 [15]. Beyond financial harm, the reputational damage from deepfakes is equally concerning. Victims often suffer from long-term erosion of trust, reputational fallout, and brand damage. In one instance, a fabricated image of an explosion near the Pentagon, generated by AI, caused a temporary dip in the Dow Jones index, highlighting how deepfakes can disrupt public confidence and financial stability [42].

These statistics emphasize the urgent need for generalized forensic methods to detect deepfake content, thereby protecting individuals, institutions, and the public at large. Several methods [17, 39, 51, 56, 60] have been developed using deep learning [60] and hybrid [39] approaches to ensure media integrity. Some approaches [51, 60] incorporate preprocessing and data augmentation techniques to enhance generalization, while others [39] leverage SoTA foundation models for robust feature extraction to ensure better generalizability.

The major drawbacks of these methods [17, 39, 51, 60] are specific to datasets [60], generative models [17, 59], or training strategies [51, 58]. Although many benchmark datasets [51] are now publicly available for evaluating a model's effectiveness and generalizability, most existing approaches [17, 39, 51, 60] consider only one or two datasets for evaluation, leaving many others unexplored. Furthermore, these methods are not assessed within a unified framework, which hinders the reproducibility of results and limits future research.

In this article, we conduct an empirical study of generalizable and explainable deepfake detection. The major contributions are listed as follows:

- We propose a unified benchmarking framework that systematically evaluates the generalization capabilities of SoTA forensic methods across benchmark datasets, generative models, and training paradigms.
- We conduct an extensive empirical study involving ten SoTA detection methods (scratch, frozen, and finetuned) and seven publicly available deepfake datasets (GAN and Diffusion) that offer a comprehensive and reproducible evaluation setup.
- We incorporate explainability techniques (confidence, ROC curves, and GradCAM) to interpret model predictions and highlight the decisions made by them.
- We provide critical insights into the strengths and limitations of current forensic methods and identify open challenges that aim to guide the development of more robust, generalizable, and explainable deepfake detection methods.

2 Empirical Study Design

This section outlines the overall design of the empirical comparative study depicted in Figure 1, covering benchmark selection, evaluation protocols, and explainability techniques. We

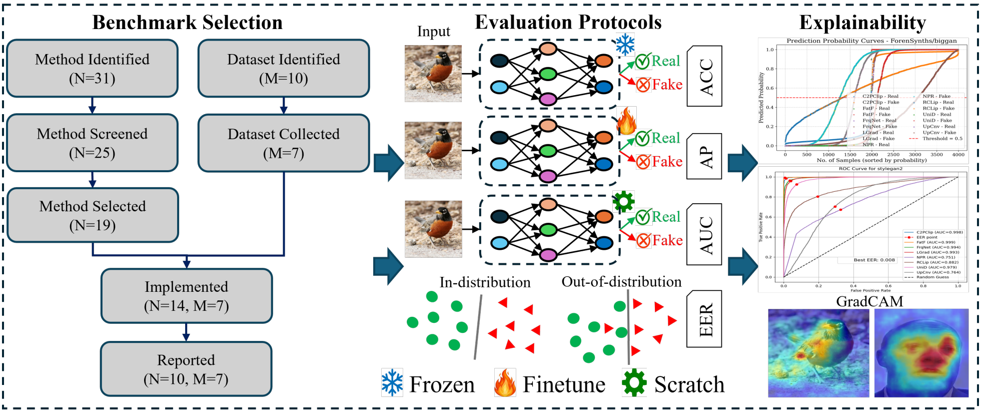


Figure 1: Overview of the empirical study: First, we perform benchmark selection, including datasets, forensic, and explainability techniques. Next, we define evaluation protocols covering frozen, fine-tuned, and from-scratch models for AI-generated image detection. Finally, we provide a comprehensive explanation based on confidence, ROC curves, and GradCAM.

Table 1: Summary of Benchmark Datasets Commonly Used in the Research Community

Name	Year	Generative Technique
ForenSyn [60]	2020	ProGAN [23], StyleGAN [24], BigGAN [4], CycleGAN [64], StarGAN [8], GauGAN [40], StyleGAN2 [25], Deepfakes [46]
ForenSynthsCh [60]	2022	CRN [6], IMLE [30], SAN [11], SIDT [6], WFR [60]
Diffusion1KStep [51]	2023	Dalle [41], DDPM [22], Guided-Diffusion [12], Improved-Diffusion [37], Mid-Journey [51]
DIRE [61]	2024	ADM [12], DDPM [22], IDDPM [37], LDM [45], PNDM [32], SDV1 [45], SDV2 [45], VQDiffusion [20]
GAN [51]	2024	AttGAN [21], BEGAN [3], CramerGAN [2], InfoMaxGAN [28], MMDGAN [29], RelGAN [38], S3GAN [34], SNGAN [35], STGAN [33]
UClipiffusion [39]	2023	Dalle [41], Glide (50_27, 100_10, 100_27) [36], Guided [12], LDM (100, 200, 200_cfg) [45]
MNW [44]	2025	Adobe, Adversarial, Amazon_v2, Aura_flow, Baidu, Bytedance_v3, Civitai_v6, Flux, Google, Hunyuandit, Hypersd, Ideogram, Kandinsky, Krea_1, Kuaishou, Luma_photon, Lumina, Meta_imagine, Midjourney, Nvidia_sana, Openai, Pixart_alpha_xl, Playgroundai, Recraft_v3, Reve_ai, Stable, Ultrapixel, Wuerstchen

identified 31 forensic methods and 10 benchmark datasets. Among the 31 forensic methods, we screened 25 and selected 19 based on venue, effectiveness, and novelty. We implemented 14 forensic methods. Similarly, we identified 10 benchmark datasets and collected 7 of them based on accessibility and representation of recent generative models. Owing to the limited generalization ability of the 4 implemented methods and their outdated nature, we ultimately reported generalization and explainability results using 10 forensic methods across 7 datasets, as listed in Table 1 and Table 2.

2.1 Datasets

This section provides an overview of SoTA benchmark datasets, including their names, release years, object categories, and generative techniques, as summarized in Table 1.

2.1.1 GAN-Based Datasets

The **ForenSyn [60]** dataset was introduced by Wang et al. to improve the generalization capability of generic deepfake detection. It comprises data from eight GAN sources, including three conditional GANs [8, 40, 64], unconditional GANs [23, 24], and a deepfake face [46] source. Most SoTA methods train their models on the ProGAN [23] training set. **GAN [51]** contains data from 9 GAN sources with varying architectural properties. These data differ from ForenSyn [60], which covers a diverse range of wild scenes. In ForenSyn [60], each sub-dataset has a random number of real and deepfake samples, while in GAN [51], each

Table 2: Summary of Benchmark Detection Methods Used in the Research Community

Name	Year	Strength	Limitation
CNND [60]	CVPR, 2020	Improved generalization by careful data augmentation	Did not explore other types of generative models such as diffusion
LGrad [49]	CVPR, 2023	Extracted gradient features using pretrained generative models	Limited to StyleGAN and ProGAN models Did not explore any diffusion models
NPR [51]	CVPR, 2024	Explored neighboring pixel-relationship	Bias towards the deepfake class
UClip [39]	CVPR, 2023	Does not require retraining of CLIP	Limited generalization to recent generative models
RClip [10]	CVPR, 2024	Requires less training data	Limited to certain generative models
FatF [31]	CVPR, 2024	Captures frequency domain artifacts	Lacks generalization to recent unseen datasets
RINE [27]	ECCV, 2024	Trainable importance estimator for encoder	Evaluated on fewer datasets
UpConv [13]	CVPR, 2020	Captured spectral features	Comparatively less effective and less generalizable
FreqNet [50]	AAAI, 2024	End-to-end frequency learning model	Does not account for other image properties
C2PClip [48]	AAAI, 2025	Caption generation and enhancement Concept injection to finetune CLIP	Limited analysis of the captions results in incomplete information

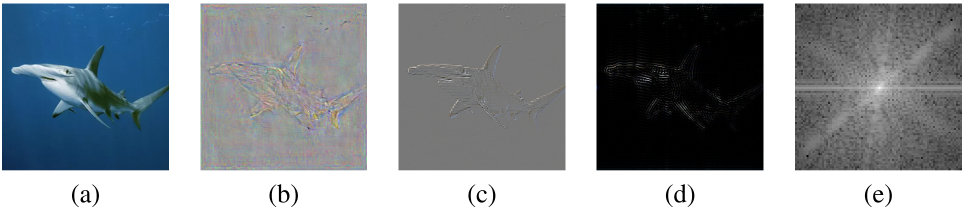

(a) (b) (c) (d) (e)

Figure 2: Intermediate representation: (a) Original, (b) LGrad (gradient) [49], (c) NPR [51], (d) FreqNet (high-frequency) [50], and (e) UpConv(spectral) [13] .

sub-dataset comprises 2K real and 2K deepfake images.

2.1.2 Diffusion-based Datasets

DIRE [61] consists of 8 diffusion-generated deepfake samples. The real images are randomly collected from ForenSyn [60] (LSUN [60] and ImageNet [47]) real classes. **Diffusion1kStep [51]** is a diffusion-family dataset containing data from five diffusion sources. All samples are generated using 1K diffusion steps. Among these, the Mid-Journey and Dalle samples were collected from social platforms. **UClipiffusion [39]** is another diffusion dataset, collected from UClip [39], encompassing four different diffusion models with varying configurations. **Microsoft Northwestern Witness (MNW) [44]** is a recent and diverse dataset encompassing 43 diffusion models, with 250 samples generated for each model. The dataset includes samples from wild scenes, faces, and real-world scenarios.

2.1.3 Other Generative Datasets

The **ForenSynthsCh [60]** dataset contains AI-generated images created using low-level vision and perceptual loss techniques, which are very challenging and often overlooked by most SoTA methods, as they worked very poorly.

2.1.4 Detection Methods

This section introduces the SoTA forensic methods used in our analysis, as summarized in Table 2, and presents their intermediate representations in Figure 2 to better illustrate the underlying concepts.

2.1.5 Scratch Trained Models

We selected four scratch-trained models [13, 49, 50, 51] to evaluate the proposed benchmark, each representing a distinct design in AI-generated image detection. UpConv [13] is a widely recognized approach that exploits spectral analysis to identify upsampling artifacts, effectively capturing intrinsic properties of both GAN- and diffusion-generated content. LGrad [49] is another influential method, which leverages a pretrained generative model to extract gradient-based features, thereby capturing subtle textural and structural cues associated with deepfakes. The nearest pixel relationship (NPR) [51] method takes a spatial perspective, focusing on the correlations among neighboring pixels to uncover artifacts introduced during the upsampling process. Finally, FreqNet [50] represents a recent advancement

Table 3: Performance evaluation on Diffusion1kStep [51] datasets (ACC/AP).

Dataset	Scratch Models				Frozen Models		Fine-Tuned Models			
	UpConv [13]	LGrad [49]	NPR [51]	FreqNet [50]	UClip [39]	RCLip [10]	RINE [27]	CNND [60]	FatF [31]	C2PClip [48]
Dalle	47.9/46.5	76.6/87.1	69.5/86.4	51.3/58.9	53.7/69.3	76.8/81.1	60.5/86.2	52.8/53.4	68.8/93.2	66.4/93.1
Ddpm	49.9/49.8	59.8/80.3	70.8/81.9	69.4/86.6	72.2/84.4	65.6/74.1	68.6/85.1	50.2/59.0	59.1/77.9	72.0/81.9
Guided-diffusion	57.6/68.3	68.5/74.8	64.3/77.6	80.3/90.2	77.5/94.5	70.0/80.1	82.2/97.7	56.4/67.7	81.8/95.7	74.4/94.6
Improved-diffusion	53.8/62.8	42.3/43.9	68.7/87.0	52.9/60.5	69.2/90.8	51.3/50.1	66.6/92.8	47.3/51.4	59.4/72.5	75.2/92.6
Midjourney	51.7/53.4	64.1/71.4	68.8/88.2	53.4/61.7	49.9/48.5	62.7/65.3	53.3/67.1	49.0/42.1	62.7/85.4	66.8/93.1
Avg.	**52.2/56.1**	**62.3/71.5**	**68.4/84.2**	**61.5/71.6**	**64.5/77.5**	**65.3/70.2**	**66.2/85.8**	**51.1/54.7**	**66.4/84.9**	**70.9/91.0**

in frequency-domain approaches, offering an end-to-end frequency-aware architecture capable of identifying nuanced spectral inconsistencies present in AI-generated images.

2.1.6 Frozen Models

Similar to scratch-trained models, we selected two frozen-based models [10, 39]. These two methods used CLIP as a frozen model to extract features for deepfake detection. First, universal deepfake detection using CLIP (UClip) [39], in which the authors adopted a pretrained CLIP model for AI-generated image detection without training CLIP. RClip [10] investigated the effectiveness of sample size in generalizing detection with CLIP features, showing that even 0.01K samples are sufficient to detect deepfake artifacts.

2.1.7 Fine-Tuned Models

This category of methods [63] fine-tunes pretrained models on AI-generated datasets to improve generalization. Examples include CNND [60], RINE [27], FatF [31], and C2PClip [48] . CNND [60] was the first to introduce a large-scale deepfake dataset, using a pretrained ResNet (trained on ImageNet) fine-tuned on this dataset. RINE [27] employs a frozen CLIP encoder with a trainable importance estimator to select key features for AI-generated image detection. Similarly, FatFormer integrates a forgery-aware adapter to capture frequency cues, while C2PClip [48] injects category-common prompts to enhance generalization.

3 Results

This section presents the experimental setups, performance evaluations and comparisons, and explainability analyses conducted in the proposed empirical study.

3.1 Experimental Setting

We configured our pipeline to evaluate all selected methods under identical environmental settings. We run all the experiments on a Linux 24.04 operating system with eight NVIDIA RTX 6000 Ada Generation GPUs (49 GB of memory on each GPU). For each method, we adopted the preprocessing, including load size, cropping, and normalization, reported in the original papers. We reported ACC, AP, AUC, and EER for a fair assessment of the methods.

3.2 Performance Evaluation and Comparisons

We extensively evaluated the performance of ten forensic methods on seven benchmark datasets, as reported in Tables 3-9. For the CNND [60] dataset, we split it into two categories because most methods tend to ignore the ForenSynthsCh [60] segments. This is because many SoTA methods fail to generalize on this dataset, resulting in poor performance. Across most datasets, UpConv [13] underperforms, while C2PClip [48] consistently achieves the highest accuracy, demonstrating strong generalization. All methods struggle on Diffusion1kStep [51], whereas UDiffusion and GAN-based datasets are easier to detect, with several models exceeding 90% accuracy. The best performance on Diffusion1kStep [51] is achieved by C2PClip [48] (ACC/AP of 70.9%/91.0%), whereas the lowest results are reported by CNND [60] (ACC/AP of 51.1%/54.7%). Methods like LGrad [49], RCLip [10], and CNND [60] show intermediate performance across all datasets, as depicted in Figure 3. Among all methods, NPR [51] achieves the highest average accuracy (91.1%) on the MNW [44] dataset, whereas most other models perform substantially worse and face the difficulty of generalizing across diverse generative sources. Notably, FreqNet [50] drops to only 1.6% accuracy, despite its strong performance on other datasets, which highlights the challenges of adapting to certain real-world or unseen data distributions.

Table 4: Performance evaluation on DIRE [61] datasets (ACC/AP).

Dataset	Scratch Models				Frozen Models		Fine-Tuned Models			
	UpConv [13]	LGrad [49]	NPR [51]	FreqNet [50]	UClip [39]	RCLip [11]	RINE [27]	CNND [60]	FatF [51]	C2PClip [48]
Adm	56.1/65.0	83.8/94.3	68.8/80.8	66.7/85.2	67.9/86.3	81.6/96.4	69.7/92.5	58.0/74.8	70.7/93.7	68.8/95.3
Ddpm	55.1/33.6	81.2/92.4	67.2/97.2	90.3/99.1	80.7/96.4	72.1/69.2	80.7/96.8	62.9/64.3	67.2/78.9	73.5/76.2
Iddpm	46.9/46.1	63.3/84.9	71.8/94.3	60.1/92.9	73.4/96.7	69.7/82.2	75.2/97.9	50.4/74.9	69.3/96.3	80.7/94.9
Ldm	63.5/67.2	98.7/99.9	74.0/99.6	97.5/100.0	50.7/86.1	95.6/100.0	56.6/98.1	53.0/75.8	97.2/100.0	97.2/99.7
Pndm	52.4/53.6	67.8/94.2	73.2/85.9	85.0/99.3	86.2/99.1	95.5/99.8	83.8/99.0	50.9/76.6	99.2/100.0	84.2/97.2
Sdv1	42.0/74.2	83.2/97.5	82.4/94.9	93.8/99.6	52.8/90.8	68.0/96.8	78.0/98.8	39.1/78.0	61.6/97.0	78.9/99.2
Sdv2	61.6/67.1	96.7/99.8	74.0/98.9	70.7/96.5	53.3/85.0	46.2/36.2	57.4/89.9	52.2/72.9	84.4/98.7	66.7/94.8
Vqdiffusion	65.3/70.4	86.1/99.0	74.0/99.6	99.9/100.0	77.8/99.0	95.6/100.0	91.4/99.9	53.9/84.7	100.0/100.0	95.8/99.7
Avg.	**55.4/59.6**	**82.6/95.3**	**73.2/93.9**	**83.0/96.6**	**67.9/92.4**	**78.0/85.1**	**74.1/96.6**	**52.6/75.2**	**81.2/95.6**	**80.7/94.6**

Table 5: Performance evaluation on ForenSynths [60] datasets (ACC/AP).

Dataset	Scratch Models				Frozen Models		Fine-Tuned Models			
	UpConv [13]	LGrad [49]	NPR [51]	FreqNet [50]	UClip [39]	RCLip [11]	RINE [27]	CNND [60]	FatF [51]	C2PClip [48]
Biggan	67.3/81.9	74.5/78.3	58.4/65.2	91.2/96.2	95.1/99.3	80.4/95.6	99.6/99.9	70.2/84.5	99.5/100.0	99.1/100.0
Cyclegan	69.7/79.3	80.1/88.3	73.8/71.3	95.5/99.6	98.3/99.8	93.5/99.5	99.3/100.0	85.2/93.5	99.4/100.0	97.3/100.0
Gaugan	59.6/74.1	68.8/73.4	53.5/49.7	92.9/98.4	99.5/100.0	91.8/97.9	99.8/100.0	78.9/89.5	99.4/100.0	99.2/100.0
Progan	53.1/78.8	98.8/99.9	58.1/71.7	99.6/100.0	99.8/100.0	84.0/99.7	100.0/100.0	100.0/100.0	99.9/100.0	100.0/100.0
Stargan	92.8/100.0	95.7/99.8	63.5/99.0	84.3/99.3	95.7/99.4	61.4/98.8	99.5/100.0	91.7/98.1	99.7/100.0	99.6/100.0
Stylegan	60.1/74.7	92.6/99.3	65.4/84.6	91.2/99.8	84.9/97.6	84.9/94.0	88.9/99.4	87.1/99.6	97.1/99.8	96.4/99.5
Stylegan2	53.8/68.6	93.6/99.2	61.7/74.8	87.3/99.5	75.0/97.9	80.8/90.2	94.5/100.0	84.4/99.1	98.8/99.9	95.6/99.9
Deepfake	53.6/53.5	58.9/81.8	49.9/52.9	92.2/97.3	68.6/81.8	53.3/72.8	80.6/97.9	53.5/89.0	93.3/98.0	93.8/98.6
Avg.	**63.8/76.4**	**82.9/90.0**	**60.5/71.2**	**91.8/98.8**	**89.6/97.0**	**78.7/93.6**	**95.3/99.7**	**81.4/94.2**	**98.4/99.7**	**97.6/99.7**

Table 6: Performance evaluation on ForenSynthsCh [60] datasets (ACC/AP).

Dataset	Scratch Models				Frozen Models		Fine-Tuned Models			
	UpConv [13]	LGrad [49]	NPR [51]	FreqNet [50]	UClip [39]	RCLip [11]	RINE [27]	CNND [60]	FatF [51]	C2PClip [48]
CRN	52.5/60.1	51.2/64.7	48.8/45.5	53.7/74.8	56.6/96.6	61.3/83.1	89.3/97.3	86.3/98.2	69.5/99.8	93.3/99.9
IMLE	51.6/62.5	51.2/70.9	48.8/50.7	53.7/69.9	69.1/98.6	66.1/83.2	90.7/99.7	86.2/98.4	69.5/99.9	93.3/99.9
SAN	50.5/48.0	42.0/41.3	58.7/68.4	89.3/93.2	56.6/78.8	76.5/88.0	68.3/94.9	50.5/70.4	68.0/81.2	64.4/84.6
SITD	85.0/97.1	47.2/39.1	51.7/53.0	72.8/72.1	62.2/63.8	70.6/91.2	90.6/97.2	90.3/97.2	81.4/97.9	95.6/98.9
WFR	64.1/84.0	57.8/58.9	51.0/49.4	50.9/96.7	87.2/97.3	71.4/90.3	97.0/99.5	86.8/94.8	88.1/98.5	94.8/99.5
Avg.	**60.7/70.3**	**49.9/55.0**	**51.8/53.4**	**64.1/81.3**	**66.3/87.0**	**69.1/87.2**	**87.2/97.7**	**80.0/91.8**	**75.3/95.5**	**88.3/96.6**

Table 7: Performance evaluation on GAN [51] datasets (ACC/AP).

Dataset	Scratch Models				Frozen Models		Fine-Tuned Models			
	UpConv [13]	LGrad [49]	NPR [51]	FreqNet [50]	UClip [39]	RCLip [11]	RINE [27]	CNND [60]	FatF [51]	C2PClip [48]
Attgan	48.5/41.9	53.1/76.6	86.4/98.0	90.3/98.5	90.8/97.0	81.3/94.9	99.2/100.0	65.8/91.4	99.3/100.0	90.4/99.8
Began	48.9/47.9	51.0/70.4	55.2/78.7	65.4/99.3	89.3/96.3	99.9/100.0	97.9/99.9	69.7/91.9	99.9/100.0	94.8/100.0
Cramergan	73.5/84.4	50.9/59.1	73.4/92.7	99.6/100.0	90.7/99.3	68.0/90.0	97.0/99.9	91.9/99.1	98.4/100.0	98.4/100.0
Infomaxgan	42.2/42.2	53.9/82.1	74.4/92.6	63.2/95.0	88.5/96.9	68.0/90.2	96.5/99.6	62.5/86.7	98.4/100.0	98.4/100.0
Mmdgan	76.1/87.0	51.1/66.5	74.0/93.5	98.0/99.9	90.6/99.2	68.0/90.1	97.0/99.9	86.4/98.2	98.4/100.0	98.4/100.0
Relgan	93.7/98.2	74.5/95.6	88.1/99.9	99.9/100.0	93.4/98.0	80.1/98.8	99.4/100.0	88.8/98.9	99.5/100.0	92.0/99.8
S3gan	96.5/99.6	73.3/75.9	73.2/82.7	88.6/94.1	94.1/98.8	85.1/99.0	98.6/99.9	69.0/80.7	99.0/100.0	99.0/100.0
Sngan	65.5/73.3	52.3/82.5	57.8/64.4	51.2/84.7	88.6/96.8	67.9/81.7	96.7/99.7	60.8/86.6	98.3/99.9	98.4/99.9
Stgan	85.7/95.9	50.5/75.7	91.4/99.1	98.0/100.0	82.8/91.6	61.5/89.8	93.7/99.1	65.2/96.5	98.8/99.8	97.6/99.6
Avg.	**70.1/74.5**	**56.7/76.1**	**74.9/89.1**	**83.8/96.8**	**89.9/97.1**	**75.5/92.7**	**97.3/99.8**	**73.3/92.2**	**98.9/100.0**	**96.4/99.9**

Table 8: Performance evaluation on UClipiffusion [39] datasets (ACC/AP).

Dataset	Scratch Models				Frozen Models		Fine-Tuned Models			
	UpConv [13]	LGrad [49]	NPR [51]	FreqNet [50]	UClip [39]	RCLip [11]	RINE [27]	CNND [60]	FatF [51]	C2PClip [48]
Dalle	55.1/65.5	83.5/92.4	53.8/69.5	97.7/99.5	87.5/97.7	89.2/99.5	95.0/99.5	56.1/71.3	98.7/99.8	98.6/99.9
Glide_50_27	58.1/67.0	85.2/92.3	54.0/80.8	86.6/95.8	79.2/96.0	87.2/96.7	92.6/99.5	62.7/84.6	94.6/99.5	95.2/99.8
Glide_100_10	59.7/69.1	83.7/91.5	54.1/81.0	88.4/96.2	78.0/95.5	87.9/97.0	90.7/99.2	61.0/82.0	94.2/99.3	96.1/99.8
Glide_100_27	54.5/60.7	81.5/89.2	53.9/80.0	84.7/95.4	78.6/95.8	87.8/97.0	88.9/99.1	60.4/80.5	94.3/99.3	95.2/99.7
Guided	57.5/68.7	70.2/75.1	58.8/67.3	62.4/67.2	70.0/88.3	85.6/96.6	76.1/96.6	62.0/77.7	76.0/91.9	69.1/94.1
Ldm_100	49.5/54.9	86.4/93.7	54.4/82.7	97.0/99.9	95.2/99.3	89.5/99.9	98.7/99.9	55.1/72.5	98.6/99.9	99.3/100.0
Ldm_200_cfg	51.3/56.7	88.2/95.4	54.3/82.9	96.9/99.8	74.2/93.2	89.3/99.7	88.2/98.7	55.2/73.0	94.8/99.2	97.2/99.8
Ldm_200	49.0/54.2	86.1/93.7	54.4/82.6	96.9/99.8	94.5/99.4	89.5/99.9	98.3/99.9	53.9/71.1	98.6/99.8	99.2/100.0
Avg.	**54.3/62.1**	**83.1/90.4**	**54.7/78.4**	**88.8/94.2**	**82.2/95.7**	**88.3/98.3**	**91.1/99.0**	**58.3/76.6**	**93.7/98.6**	**93.8/99.1**

3.3 Explainability of Model Predictions

For a better explanation of model predictions, we visualized GradCAM, confidence, and ROC curves, as depicted in Figures 4, 5, and 6. GradCAM highlights the regions that each model focuses on to distinguish between real and fake samples. As shown in Figure 4, each method focuses on different regions to determine whether a sample is real. For example, LGrad [49], FreqNet [50], and C2PClip [48] primarily target the background, while others attend to random regions when making their decisions.

Table 9: Performance evaluation on MNW [44] datasets (ACC/AP).

Dataset	**Scratch Models**				**Frozen Models**		**Fine-Tuned Models**			
	UpConv [13]	**LGrad [illegible]**	**NPR [51]**	**FreqNet [illegible]**	**UClip [illegible]**	**RCLip [illegible]**	**RINE [illegible]**	**CNND [illegible]**	**FatF [illegible]**	**C2PClip [illegible]**
Adobe	41.6/–	56.5/–	97.6/–	1.2/–	28.9/–	4.7/–	38.5/–	3.1/–	16.7/–	31.3/–
Adversarial_images	15.2/–	23.6/–	89.2/–	2.4/–	10.4/–	5.2/–	8.8/–	1.2/–	6.4/–	6.4/–
Amazon_titan_v2	10.8/–	72.0/–	100.0/–	0.0/–	20.0/–	29.6/–	27.6/–	3.2/–	23.6/–	26.4/–
Aura_flow	18.4/–	56.8/–	100.0/–	0.8/–	2.4/–	4.8/–	3.6/–	0.0/–	12.0/–	47.2/–
Baidu	10.4/–	13.2/–	85.2/–	0.8/–	10.4/–	1.2/–	6.8/–	0.8/–	0.4/–	4.4/–
Bytedance	31.6/–	49.6/–	99.2/–	0.0/–	0.0/–	1.2/–	0.4/–	0.4/–	0.0/–	0.0/–
Civitai_v6	3.6/–	80.0/–	98.0/–	0.0/–	4.0/–	12.0/–	4.8/–	0.8/–	21.6/–	28.0/–
Flux	12.8/–	35.2/–	88.0/–	15.1/–	3.7/–	2.1/–	3.1/–	2.4/–	2.9/–	3.1/–
Google	5.8/–	13.8/–	70.8/–	2.2/–	4.8/–	1.4/–	1.0/–	2.0/–	0.0/–	0.2/–
Hunyuandit	32.0/–	1.6/–	94.0/–	0.0/–	7.2/–	1.2/–	4.4/–	0.8/–	0.4/–	11.2/–
Hypersd	19.6/–	3.6/–	68.0/–	0.6/–	3.2/–	0.0/–	0.8/–	1.4/–	0.0/–	3.4/–
Ideogram	6.8/–	78.4/–	98.0/–	0.4/–	2.0/–	0.0/–	2.0/–	4.0/–	14.8/–	1.2/–
Kandinsky	17.2/–	1.6/–	88.4/–	1.2/–	11.2/–	0.0/–	4.0/–	0.0/–	0.0/–	6.0/–
Krea_1	6.0/–	90.4/–	98.8/–	0.0/–	4.8/–	0.0/–	5.6/–	6.0/–	7.2/–	1.2/–
Kuaishou_kolors	8.0/–	2.8/–	85.2/–	1.6/–	3.2/–	0.4/–	0.8/–	0.4/–	0.0/–	9.6/–
Luma_photon	97.6/–	84.0/–	99.2/–	1.2/–	26.8/–	5.2/–	45.2/–	14.4/–	41.6/–	16.8/–
Lumina	20.4/–	82.4/–	99.2/–	0.4/–	15.6/–	13.2/–	30.4/–	8.4/–	27.6/–	10.8/–
Meta_imagine	4.0/–	15.2/–	94.0/–	1.2/–	14.8/–	2.8/–	12.0/–	5.6/–	0.0/–	12.4/–
Midjourney	30.4/–	30.8/–	78.9/–	0.8/–	5.7/–	0.0/–	8.8/–	10.9/–	6.1/–	7.3/–
Nvidia_sana	47.6/–	22.4/–	95.6/–	1.2/–	60.4/–	30.4/–	60.4/–	0.4/–	49.6/–	54.4/–
Openai	7.3/–	37.9/–	86.5/–	2.4/–	20.3/–	2.0/–	20.7/–	8.0/–	2.9/–	14.5/–
Pixart_alpha_xl	32.0/–	3.6/–	82.0/–	0.8/–	2.0/–	0.8/–	0.8/–	1.2/–	0.0/–	12.0/–
Playgroundai	22.0/–	38.8/–	80.2/–	0.0/–	6.2/–	1.0/–	9.2/–	3.2/–	27.4/–	10.2/–
Recraft_v3	46.0/–	66.0/–	93.6/–	0.0/–	12.4/–	0.4/–	7.2/–	0.0/–	6.8/–	1.6/–
Reve_ai	13.2/–	72.0/–	99.2/–	0.8/–	2.8/–	0.0/–	7.2/–	4.0/–	13.2/–	1.2/–
Stable_diffusion	14.2/–	41.1/–	89.6/–	2.0/–	14.3/–	9.2/–	13.7/–	3.9/–	17.1/–	17.2/–
Ultrapixel	11.2/–	84.0/–	100.0/–	5.6/–	4.0/–	0.0/–	4.0/–	44.4/–	38.8/–	0.8/–
Wuerstchen	8.4/–	4.0/–	93.6/–	1.2/–	22.4/–	1.2/–	30.4/–	5.6/–	0.8/–	17.2/–
Avg.	**21.2/–**	**41.5/–**	**91.1/–**	**1.6/–**	**11.6/–**	**4.6/–**	**12.9/–**	**4.9/–**	**12.1/–**	**12.7/–**

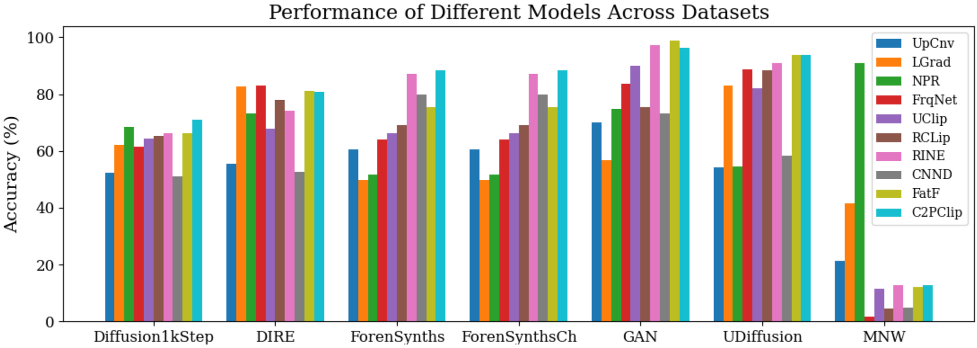


Figure 3: Summary of all forensic methods on all benchmark datasets.

In contrast, the confidence curve represents the prediction probabilities of each model for the real and fake classes to make it clear how confident a model is in predicting real as real and fake as fake. As shown in Figure 5, in most cases, NPR [51] is biased towards the fake class, while UpConv [13] tends to favor the real class. Similar to the confidence curve, the ROC curve illustrates the trade-off between the true positive rate and false positive rate across different thresholds to provide a comprehensive view of each model's discriminative ability.

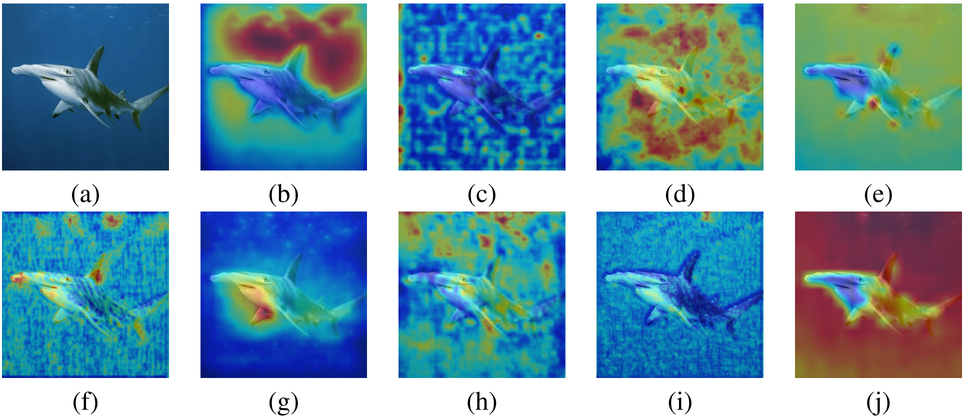
4 Discussions and Future Research Directions

This section outlines the discussions and future research directions of our findings.

4.1 Discussions

Inconsistent experimental settings: While most methods employ the same training set, variations in their basic experimental configurations lead to inconsistencies across SoTA methods, thereby hindering the reproducibility of results reported in the paper.

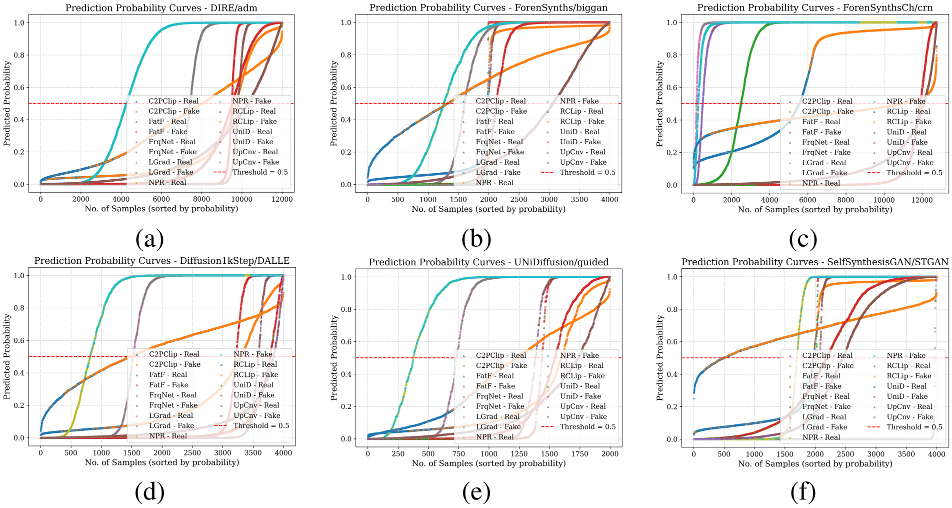
Lack of generalization: Although most methods claim to generalize to unseen generative models, they struggle with unseen samples, as shown in Tables 3–9, particularly for the MNW [44] dataset in Table 9 while varying the generative models.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 4: GradCAM explanation: (a) Original, (b) LGrad [49], (c) NPR [51], (d) FreqNet [50], (e) UClip [39], (f) RClip [10], (g) RINE [27], (h) CNND [60], (i) FatF [31], and (j) C2PClip [48].



(a) (b) (c)

(d) (e) (f)

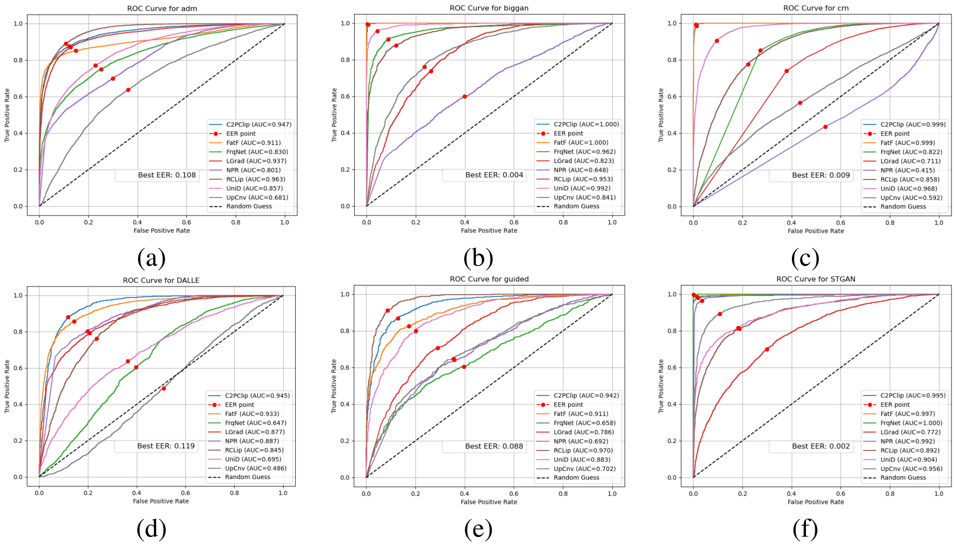
Figure 5: Confidence of fake prediction by each model on six datasets: (a) Adm, (b) BigGAN, (c) CRN, (d) Dalle, (e) Guided, and (f) StGAN.

Biases in decision-making: In many cases, the methods exhibit bias toward either the real or deepfake class. As shown in Table 9, most methods fail to detect MNW [44] samples, whereas NPR [51] achieves 91% ACC. Our analysis of the confidence curve reveals that NPR [51] is biased toward fakes, as shown in Figure 5, which enables it to detect the MNW [44] dataset. In contrast, UpConv [13] is biased toward real class (Figure 5).

Restricted preprocessing: Most methods rely on predefined preprocessing pipelines tailored to specific datasets; for instance, NPR [51], RINE [27], and FreqNet [50] omit cropping for certain datasets, while applying it to others.

Vulnerability to AFs: A few studies [60] have evaluated robustness against conventional AFs, such as noise and compression. However, none have considered AFs based on GANs [55], diffusion models [61], or optimization-based anti-forensic (AF) attacks.

Lack of explainability: Most methods lack explainability of their prediction to provide insight into model behavior to make it difficult to understand why a particular decision was made and limiting trust, accountability, and the ability to improve the model effectively.



(a) (b) (c)

(d) (e) (f)

Figure 6: ROC curve of each model corresponding to prediction confidence: (a) Adm, (b) BigGAN, (c) CRN, (d) Dalle, (e) Guided, and (f) StGAN.

4.2 Future Research Directions

Standardized framework: Our findings suggest developing unified training, preprocessing, and evaluation protocols to ensure fair comparisons and reproducibility of the results.
Improved generalization: To improve generalization to GANs and diffusion, our study suggests domain-agnostic features using meta-learning or multi-domain training.
Bias reduction: To better generalize across real and fake classes, our framework recommends balanced objectives and debiasing techniques to avoid skew toward one class.
Preprocessing robustness: Additionally, the experimental results encourage building models that work reliably across varied or minimal preprocessing and AF conditions.
Explainability and trust: Future research should focus on enhancing model interpretability through explainable AI techniques, such as attention visualization, causal reasoning, rule-based representations, and large language model–driven report generation, to improve trust and usability in real-world forensic applications.

5 Conclusions

In this study, we evaluated SoTA forensic methods under unified configurations across multiple benchmark datasets to reveal their strengths and limitations. Our empirical analysis highlighted key challenges, including inconsistent experimental settings, limited generalization to unseen generative models, biases in decision-making, and dependence on dataset-specific preprocessing. By systematically benchmarking ten SoTA methods across seven datasets, we provided insights into their generalization and applicability in real-world scenarios.
Furthermore, we proposed future research directions, including the development of standardized frameworks, improved generalization through domain-agnostic feature learning, bias mitigation strategies, and preprocessing-robust model design. Overall, this study serves as a comprehensive guide for the research community to inspire the development of more robust, generalizable, and explainable approaches for detecting AI-generated media. **[The code, model weights, and datasets will be released upon acceptance of the paper.]**

References

[1] Shruti Agarwal and Hany Farid. Detecting deepfake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[2] Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Remi Munos. The cramer distance as a solution to biased wasserstein gradients. 2017. In *URL https://openreview. net/forum*.

[3] David Berthelot, Thomas Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. arxiv 2017. *arXiv preprint arXiv:1703.10717*, 2017.

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[5] CFO.com. Most companies have experienced financial loss due to a deepfake, 2024. URL https://www.cfo.com/news/most-companies-have-experienced-financial-loss-due-to-a-deepf 732094/. [Online; accessed 2024].

[6] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017.

[7] Robert Chesney and Danielle Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(1):175–200, 2019.

[8] 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 conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[9] Kim-Kwang Raymond Choo. The dangers of fake content in the age of iot and ai. *Computer Fraud & Security*, 2019(5):10–13, 2019.

[10] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip (2023). *arXiv preprint arXiv:2312.00195*, 2024.

[11] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11065–11074, 2019.

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[13] 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.

[14] Eftsure. Statistics: Deepfake fraud in 2024, 2024. URL https://www.eftsure.com/statistics/deepfake-statistics/. [Online; accessed 2024].

[15] Eftsure. Deepfake-related fraud forecast to hit $40b by 2027, 2025. URL https://www.eftsure.com/statistics/deepfake-statistics/?utm_source=chatgpt.com. [Online; accessed 2025].

[16] eSecurityPlanet.com. AI Deepfakes Surge: $200 Million Lost, 2025. URL https://www.esecurityplanet.com/news/ai-deepfakes-surge-200-million-lost/. [Online; accessed 2025].

[17] Joshua Frank, Thorsten Eisenhofer, Lea Schönherr, Andreas 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.

[18] Jason Fridman, John Brown, and Can Mericli. Synthetic video attacks on autonomous driving systems using gans. *arXiv preprint arXiv:2001.03667*, 2020.

[19] Globe Newswire. Deepfake fraud costs the financial sector an average of $600,000 per company, 2024. URL https://www.businesswire.com/news/home/20241031656724/en/Deepfake-Fraud-Costs. [Online; accessed 2024].

[20] 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.

[21] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE transactions on image processing*, 28(11):5464–5478, 2019.

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[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.

[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.

[26] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018.

[27] Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision*, pages 394–411. Springer, 2024.

[28] Kwot Sin Lee, Ngoc-Trung Tran, and Ngai-Man Cheung. Infomax-gan: Improved adversarial image generation via information maximization and contrastive learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3942–3952, 2021.

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

[30] Ke Li, Tianhao Zhang, and Jitendra Malik. Diverse image synthesis from semantic layouts via conditional imle. 2019 ieee. In *CVF International Conference on Computer Vision (ICCV)*, pages 4219–4228, 2019.

[31] 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.

[32] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022.

[33] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3673–3682, 2019.

[34] Mario Lučić, Michael Tschannen, Marvin Ritter, Xiaohua Zhai, Olivier Bachem, and Sylvain Gelly. High-fidelity image generation with fewer labels. In *International conference on machine learning*, pages 4183–4192. PMLR, 2019.

[35] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[36] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021.

[37] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

[38] Weili Nie, Nina Narodytska, and Ankit Patel. Relgan: Relational generative adversarial networks for text generation. In *International conference on learning representations*, 2018.

[39] 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.

[40] 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.

[41] 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.

[42] Reality Defender. Understanding the hidden costs of deepfake fraud in finance, 2023. URL https://www.realitydefender.com/insights/understanding-the-hidden-costs-of-deepfake-fraud-in-finance. [Online; accessed 2023].

[43] Reddit r/SocialEngineering. Deepfake-related crypto scams statistics, 2025. URL https://www.reddit.com/r/SocialEngineering/comments/1hwxvhp/how_are_scammers_using_5_deepfakes_to_steal/. [Online; accessed 2025].

[44] Thomas Roca, Marco Postiglione, Chongyang Gao, Isabel Gortner, Zuzanna Wojciak, Pengce Wang, Masah Alimardani, Shirin Anlen, Kevin White, Juan Lavista, Sarit Kraus, Sam Gregory, and V.S. Subrahmanian. Introducing the mnw benchmark for ai forensics. https://github.com/nsail-lab/MNW, 2025. PDF.

[45] 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.

[46] 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 international conference on computer vision*, pages 1–11, 2019.

[47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115 (3):211–252, 2015.

[48] Tan, Tao Chuangchuang, Liu Renshuai, Gu Huan, Wu Guanghua, Wei Baoyuan, Zhao nad Yao, and Yunchao. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7184–7192, 2025.

[49] 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.

[50] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5052–5060, 2024.

[51] 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.

[52] Nusrat Tasnim and Joong-Hwan Baek. Deep learning-based human action recognition with key-frames sampling using ranking methods. *Applied Sciences*, 12(9):4165, 2022.

[53] Nusrat Tasnim and Joong-Hwan Baek. Dynamic edge convolutional neural network for skeleton-based human action recognition. *Sensors*, 23(2):778, 2023.

[54] Nusrat Tasnim, Md Mahbubul Islam, and Joong-Hwan Baek. Deep learning-based action recognition using 3d skeleton joints information. *Inventions*, 5(3):49, 2020.

[55] Kutub Uddin, Yoonmo Yang, and Byung Tae Oh. Anti-forensic against double jpeg compression detection using adversarial generative network. *In Proceedings of the Korean Society of Broadcast Engineers Conference*, pages 58–60, 2019.

[56] Kutub Uddin, Yoonmo Yang, Tae Hyun Jeong, and Byung Tae Oh. A robust open-set multi-instance learning for defending adversarial attacks in digital image. *IEEE Transactions on Information Forensics and Security*, 19:2098–2111, 2023.

[57] Kutub Uddin, Tae Hyun Jeong, and Byung Tae Oh. Counter-act against gan-based attacks: A collaborative learning approach for anti-forensic detection. *Applied Soft Computing*, 153:111287, 2024.

[58] Kutub Uddin, Awais Khan, Muhammad Umar Farooq, and Khalid Malik. Shield: A secure and highly enhanced integrated learning for robust deepfake detection against adversarial attacks. *arXiv preprint arXiv:2507.13170*, 2025.

[59] Kutub Uddin, Nusrat Tasnim, Muhammad Saad Saeed, and Khalid Mahmood Malik. Guard: Generative unmasking and adversarial-resistant deepfake detection using multi-model knowledge distillation. *Authorea Preprints*, 2025.

[60] 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.

[61] 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.

[62] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[63] Zhen Yi, Qiang Liu, Yuan Zhang, and Li Tan. Deep learning based face recognition: A survey. *IEEE Access*, 7:106395–106413, 2019.

[64] 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 international conference on computer vision*, pages 2223–2232, 2017.