RealNet: Efficient and Unsupervised Detection of AI-Generated Images via Real-Only Representation Learning

Shuaibo Li[1,2*], Laixin Zhang[1*], Wei Ma[1†], Jianwei Guo[3], Shibiao Xu[4], Zhijie Qiu[2], Hongbin Zha[5,6]

[1]College of Computer Science, Beijing University of Technology
[2]The Hong Kong University of Science and Technology (Guangzhou)
[3]School of Artificial Intelligence, Beijing Normal University
[4]School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[5]Anqing Normal University
[6]School of IST, Peking University
sli270@connect.hkust-gz.edu.cn, mawei@bjut.edu.cn

Abstract

Detecting AI-generated images remains a persistent challenge, as existing detectors often struggle to generalize to forgeries produced by previously unseen generative models. This generalization gap mainly stems from entanglement with semantic content and overfitting to model-specific artifacts. Moreover, many state-of-the-art methods rely on large pre-trained backbones or computationally intensive pipelines, which limit their applicability in real-world, resource-constrained environments. We propose RealNet, a lightweight and unsupervised framework that constructs a disentangled, forgery-aware representation space using only real images. RealNet first extracts semantic-agnostic representations through a dual adversarial denoising mechanism, producing compact features with low intra-class variance. These representations are then perturbed in feature space to generate pseudo-negative samples, which are combined with the original real features to train a lightweight discriminator, enabling robust detection without any dependence on synthetic images during training. Comprehensive evaluations across GAN, diffusion, and emerging VAR-based paradigms demonstrate that RealNet achieves superior cross-model generalization and robustness. RealNet surpasses previous state-of-the-art approaches by 4.51% in accuracy and 3.93% in average precision, while maintaining significantly lower computational cost. Furthermore, we introduce a medically relevant synthetic image dataset and show RealNet remains effective under severe distribution shifts, highlighting its potential for deployment in high-stakes real-world scenarios. Together, these advantages position RealNet as a practical, scalable and socially impactful solution for robust AI-generated image detection.

Introduction

In recent years, AI generative technologies have progressed at an unprecedented pace, giving rise to a succession of powerful paradigms such as generative adversarial networks (GANs) (Karras et al. 2018; Brock, Donahue, and Simonyan 2019; Zhu et al. 2017; Choi et al. 2018), diffusion models (Dhariwal and Nichol 2021; Nichol et al. 2022; Midjourney 2023; Gu et al. 2022), and, most recently, visual autoregressive modeling (VAR) (Ren et al. 2024; Han et al. 2024; Chen et al. 2024b; Yao et al. 2024; Tian et al. 2024). These models are capable of synthesizing photorealistic images with remarkable fidelity and diversity, driving substantial advances across art, entertainment, and scientific applications. Yet as synthetic content becomes increasingly indistinguishable from authentic imagery, the associated societal risks continue to escalate: highly realistic fake images can exacerbate privacy breaches, amplify misinformation at scale, and undermine public trust in digital media, ultimately affecting individual reputations, social stability, and safety-critical decision making.

This rapidly evolving landscape demands detection methods that are not only highly reliable but also practically deployable at scale. A persistent and fundamental challenge, however, is generalization: while existing detectors often perform well on forgeries generated by models seen during training, their accuracy deteriorates markedly when confronted with content produced by previously unseen or newly emerging architectures, such as VAR models (Wang et al. 2020; Chen et al. 2022b; Ren et al. 2024; Han et al. 2024). This vulnerability has become increasingly pronounced as generative paradigms evolve at an accelerating pace, exposing the limitations of detectors that rely on paradigm-specific artifacts or training distributions and ultimately restricting their long-term applicability in real-world scenarios.

Recent research has attempted to improve generalization by learning forgery-aware representation and suppressing semantic information (Wang et al. 2023; Ojha, Li, and Lee 2023; Qu et al. 2024b), often exploiting spatial (Li et al. 2021, 2024; Qu et al. 2025) or frequency-domain artifacts (Zhang, Karaman, and Chang 2019a; Durall, Keuper, and Keuper 2020; Qu et al. 2023). However, these approaches typically rely on labeled synthetic forgeries or

*These authors contributed equally.
†Corresponding author.

patterns unique to known generative models, which inherently tie them to model-specific biases and limit their robustness to paradigm shifts (Zhang et al. 2025). More recent attempts leverage generic pre-trained representations, such as CLIP-ViT (Ojha, Li, and Lee 2023; Liu et al. 2024), to improve transferability, but such features remain highly entangled with semantic content (He, Chen, and Ho 2024), making them unreliable under unseen or emerging generative architectures. At the same time, detector designs increasingly depend on large neural backbones or heavyweight pretrained models, resulting in substantial computational overhead. This makes broad, efficient deployment difficult in real-world scenarios, particularly in resource-constrained or edge environments, further restricting the practical impact of current detectors. These limitations collectively underscore the need for a detection framework that is both more generalizable and more computationally efficient.

Beyond the challenges posed by rapidly evolving generative models, high-risk domains such as medical imaging face increasing yet under-recognized threats from AI-generated content. Recent advances have enabled the synthesis of highly realistic medical images that can assist training but may also mislead even experienced clinicians (Prezja et al. 2022; Li et al. 2025; Hao et al. 2025), raising concerns for clinical trust and safety-critical decision making. Moreover, the substantial distribution gap between medical and natural images further underscores the need for detection methods that remain domain-agnostic, efficient, and robust even when model-generated forgeries are unavailable during training. These observations highlight the need for a different detection paradigm capable of meeting these requirements.

To address the intertwined challenges of robustness, efficiency, and generalization without relying on synthetic training forgeries, we propose RealNet. RealNet is an unsupervised and lightweight detection framework that learns a disentangled and forgery-aware representation space directly from real images. As illustrated in Figure 1, RealNet first performs real feature encoding. An input image is processed by a Real Pattern Extractor, which is frozen during training and produces a semantic-agnostic real representation (SARR). A Feature Transformer further refines this representation through frequency-domain transformation and adaptive normalization, yielding a compact real representation (CRR) that suppresses semantic content and generator-specific biases. Based on CRR, a Feature Perturbator injects controlled Gaussian noise into the compact feature space to form pseudo-fake representations (PFR). These perturbed representations serve as training-only negative samples that capture model-agnostic forgery patterns. A lightweight discriminator then learns to separate CRR from PFR, establishing a stable decision boundary that enables robust forgery-aware representation learning without the need for synthetic images. Through this integrated design, RealNet offers a unified and computationally efficient approach for deriving forgery-sensitive representations solely from real images.

To facilitate comprehensive evaluation and to support research in high-impact domains, we additionally construct and release two complementary datasets. The first contains synthetic images produced by several advanced VAR-based models, which enables broad and rigorous benchmarking across emerging generative paradigms. The second contains medical forgeries created by state-of-the-art medical image synthesis methods, reflecting realistic and safety-critical distribution shifts. These datasets offer valuable resources for assessing robustness under both challenging generative processes and important real-world application scenarios. Our main contributions are summarized as follows:

- We propose RealNet, an unsupervised and lightweight detection framework that learns a disentangled forgery-aware representation space directly from real images, effectively reducing both semantic interference and generator-specific biases.
- We develop a positive-compression mechanism and a pseudo-negative sample generation strategy that enable a compact and robust decision boundary without any reliance on synthetic image supervision, thereby supporting strong generalization and improved domain adaptation.
- RealNet achieves state-of-the-art detection performance, cross-model generalization, and robustness across a wide range of generative paradigms, including challenging VAR-based models, while maintaining computational efficiency suitable for large-scale and resource-constrained deployment.
- We construct two datasets that broaden the range of evaluation settings, covering both emerging generative paradigms and sensitive application domains. These datasets enable evaluation under diverse generative processes and distribution shifts.

Related Work

Generative Frameworks in Vision Recent advances in generative modeling have led to a diverse set of architectures capable of synthesizing high-fidelity images. GANs (Karras et al. 2018; Brock, Donahue, and Simonyan 2019) learn through adversarial training, while normalizing flows (Kingma and Dhariwal 2018; Dao et al. 2023) and autoregressive models (Razavi, van den Oord, and Vinyals 2019) provide explicit likelihood formulations. Diffusion models construct a denoising process parameterized by variational or score-based objectives, gradually mapping noise to data. Visual autoregressive modeling (VAR) (Tian et al. 2024) has recently emerged as a strong alternative, generating images in a hierarchical coarse-to-fine manner that captures both global structure and fine textures. Models such as Infinity (Han et al. 2024) and FlowAR (Ren et al. 2024) demonstrate the high realism achievable under this paradigm, often surpassing diffusion models in detail and complexity. However, the rapid diversification of generative mechanisms increases the difficulty of forgery detection, as artifacts vary widely across paradigms. This underscores the need for detectors that generalize across unseen architectures without relying on access to synthetic training data.

AI-Generated Image Detection A variety of detectors have been proposed to counter increasingly powerful gen-

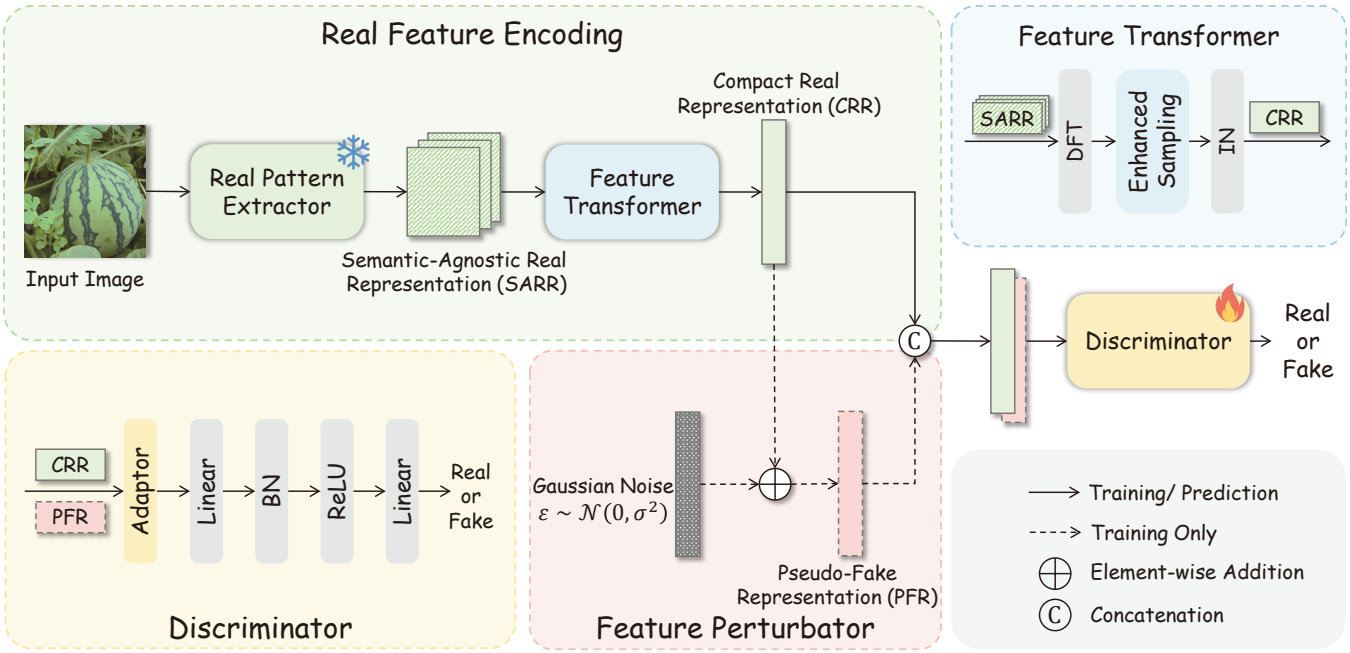

Figure 1: **Overview of RealNet**. An input image is first encoded by a frozen Real Pattern Extractor (RPE) to obtain a semantic-agnostic real representation (SARR). The Feature Transformer refines and compresses SARR into a compact real representation (CRR). During training, the Feature Perturbator injects controlled Gaussian noise into CRR to generate pseudo-fake representations (PFR), which serve as negative samples. A lightweight discriminator then learns to distinguish CRR from PFR, enabling forgery-aware representation learning without synthetic images.

erative models, which can be broadly grouped into spatial-artifact, frequency-artifact, and pre-trained-feature based methods. Spatial methods exploit architecture- or pipeline-specific traces in the image domain: for example, (Yu, Davis, and Fritz 2019) shows that each GAN leaves distinct fingerprints tied to its network structure and parameters, while (Zhao et al. 2021) formulates detection as fine-grained classification with a multi-attention network. Frequency-based approaches (Zhang, Karaman, and Chang 2019a; Dzanic, Shah, and Witherden 2020; Qu et al. 2024a) instead operate in the spectral domain, modeling discrepancies in spectrum statistics and high-frequency attenuation between real and synthetic images. More recent pre-trained-based methods build on frozen large-scale backbones with lightweight classification heads (Ojha, Li, and Lee 2023; Tan et al. 2023), which help mitigate overfitting to specific training sets and improve transferability. For instance, (Ojha, Li, and Lee 2023) leverages nearest-neighbor and linear probing on CLIP-ViT features, and (Tan et al. 2023) transforms images into gradient representations using a pre-trained CNN to capture GAN-induced artifacts.

Another complementary line of work avoids relying on fake samples for training or is entirely training-free (Ricker, Lukovnikov, and Fischer 2024; Zhang, Karaman, and Chang 2019b; Jeong et al. 2022). Methods such as (Zhang, Karaman, and Chang 2019b; Jeong et al. 2022) generate hard negative samples using predefined perturbation patterns, but their simulated artifacts are mainly effective for GAN-based models. AEROBLADE (Ricker, Lukovnikov, and Fischer 2024) exploits the observation that latent diffusion autoencoders reconstruct synthetic images more faithfully than real ones, enabling training-free detection but restricting applicability to latent diffusion outputs. In contrast, RealNet learns a disentangled representation space directly from real images in an unsupervised manner. The resulting semantic-agnostic and model-agnostic features allow our detector to avoid overfitting to forgery-irrelevant content or generator-specific patterns, leading to stronger cross-paradigm generalization.

Method

As illustrated in Figure 1, RealNet consists of four modules: a Real Pattern Extractor (RPE), a Feature Transformer, a Feature Perturbator, and a lightweight discriminator. Training is performed in two stages. In the first stage, we pre-train the RPE on real noisy–clean image pairs using a dual adversarial denoising framework and then freeze its parameters. In the second stage, only real images are used: the frozen RPE extracts a semantic-agnostic real representation (SARR), the Feature Transformer compresses SARR into a low-dimensional compact real representation (CRR), the Feature Perturbator perturbs CRR to synthesize pseudo-fake features, and the discriminator learns a decision boundary that separates CRR from its perturbed counterparts. The Feature Perturbator is used only during training and is removed at inference. We next present each component of RealNet and the overall learning pipeline in detail.

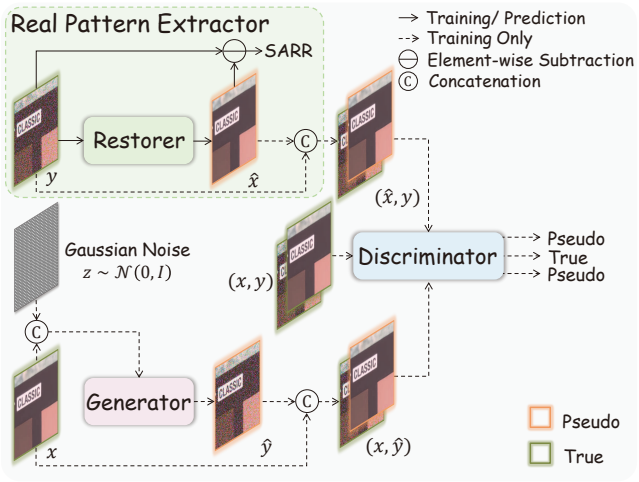

Figure 2: **Overview of the Real Pattern Extractor (RPE).** A restorer and a generator form a dual adversarial framework that constructs pseudo clean-noisy pairs from two complementary directions, while a discriminator aligns them with real image pairs. After training, the restorer is frozen and used to derive SARR through residual reconstruction.

Real Pattern Extractor

The Real Pattern Extractor (RPE) aims to suppress forgery-irrelevant content and capture the stable statistical patterns of real images, which we refer to as the Semantic-Agnostic Real Representation (SARR). For RealNet to function effectively, SARR must be compact so that real images with different semantics occupy a low-variance region. At the same time, it must be sensitive to noise so that small perturbations to the input produce noticeable changes in the representation and can serve as useful pseudo-negative samples. Prior studies (Marra et al. 2019; Wang et al. 2023) have shown that reconstruction residuals and noise-related cues are informative for distinguishing real and synthetic images. Motivated by this, we obtain SARR through an adversarial denoising and restoration framework that learns real-image noise statistics while reducing semantic influence.

To obtain such a representation, we adopt a dual adversarial denoising framework proposed by (Yue et al. 2020; LI et al. 2017). As shown in Figure 2, RPE consists of three components: a denoising restorer $R(\cdot)$, a noise-generating generator $G(\cdot)$, and a discriminator $D(\cdot)$. The restorer branch takes a real noisy image $\boldsymbol{y}$ as input and predicts a pseudo-clean image $\hat{\boldsymbol{x}} = R(\boldsymbol{y})$, forming a pseudo clean–noisy pair $(\hat{\boldsymbol{x}}, \boldsymbol{y})$. From a probabilistic perspective, with $p(\cdot)$ denoting the data distributions, this branch can be written as:

$$\boldsymbol{y} \sim p(\boldsymbol{y}), \hat{\boldsymbol{x}} = R(\boldsymbol{y}) \rightarrow (\hat{\boldsymbol{x}}, \boldsymbol{y}). \tag{1}$$

In parallel, the generator branch learns to synthesize realistic noise. It takes a real clean image $\boldsymbol{x}$ together with a latent variable $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I})$ and produces a pseudo-noisy image $\hat{\boldsymbol{y}} = G(\boldsymbol{x}, \boldsymbol{z})$, yielding another pseudo clean–noisy pair $(\boldsymbol{x}, \hat{\boldsymbol{y}})$:

$$\boldsymbol{z} \sim p(\boldsymbol{z}), \boldsymbol{x} \sim p(\boldsymbol{x}), \hat{\boldsymbol{y}} = G(\boldsymbol{x}, \boldsymbol{z}) \rightarrow (\boldsymbol{x}, \hat{\boldsymbol{y}}). \tag{2}$$

Together with real clean–noisy pairs $(\boldsymbol{x}, \boldsymbol{y})$, these two types of pseudo pairs are fed into the discriminator D, which is trained to distinguish real from synthetic pairs. Following (Liu et al. 2021), R, G, and D are jointly optimized with a dual adversarial objective:

$$\begin{aligned} \min_{R,G} \max_{D} \mathcal{L}_{\text{GAN}}(R, G, D) + \lambda_1 \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_1 \\ + \lambda_2 \|\mathcal{F}(\hat{\boldsymbol{y}} - \boldsymbol{x}) - \mathcal{F}(\boldsymbol{y} - \boldsymbol{x})\|_1, \end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{GAN}}$ is the dual adversarial loss from Triple-GAN (LI et al. 2017), $\mathcal{F}(\cdot)$ denotes a Gaussian filtering operator used to extract noise statistics, and λ_1, λ_2 are balance factors.

This adversarial training is performed on real noisy–clean image pairs from standard denoising datasets (Abdelhamed, Lin, and Brown 2018; Anaya and Barbu 2018; Xu et al. 2018). After convergence, we discard the generator and discriminator and freeze the restorer R as the Real Pattern Extractor. For any real image I_r, we reconstruct it with the pre-trained restorer F_ϕ and take the reconstruction residual as SARR:

$$f_r = F_\phi(I_r) - I_r. \tag{4}$$

Because F_ϕ has been adversarially trained to model real-image noise, the residual $f_r \in \mathbb{R}^{256\times256\times3}$ captures real-image noise patterns that are largely decoupled from semantic content yet highly sensitive to perturbations. These semantic-agnostic residuals form the input to the subsequent Feature Transformer in RealNet. Both the restorer and the generator adopt a UNet (Ronneberger, Fischer, and Brox 2015)-style architecture with residual modules, while the discriminator consists of five convolutional layers followed by a fully connected layer that aggregates pair-wise information.

Feature Transformer

The Feature Transformer compresses the SARR into a compact representation that emphasizes authenticity-related statistics while reducing semantic variation. Operating in a low dimensional space benefits RealNet by enlarging the distinction between real and perturbed features. Prior studies (Frank et al. 2020; Bi et al. 2023) also indicate that compact frequency-based embeddings help highlight subtle signal differences. As shown in Figure 1, we first apply a Discrete Fourier Transform to the SARR f_r to obtain its frequency domain representation:

$$d(u, v) = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f_r(x, y) \cdot e^{-j2\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)}. \tag{5}$$

We compute the amplitude spectrum $A_s \in \mathbb{R}^{256\times256\times3}$ and fuse the three channels using the luminosity method to produce a single channel map. To enhance discriminative frequencies, amplitudes below the mean are suppressed and the remaining values are squared, resulting in a refined spectrum $A_s{}' \in \mathbb{R}^{256\times256\times1}$. Periodic sampling and instance

normalization are then applied to $A_s{}'$, yielding the Compact Real Representation $f_{cr} \in \mathbb{R}^{64}$. This mapping is expressed as

$$f_{cr} = T_\theta(f_r), \tag{6}$$

where T_θ denotes the Feature Transformer. The resulting feature space preserves real image noise characteristics while remaining compact enough to support robust perturbation based pseudo-fake generation.

Feature Perturbator

The Feature Perturbator is designed to construct informative pseudo-negative samples directly in the learned compact space, allowing RealNet to approximate forgery-oriented variations without accessing synthetic images. Since the compact real representation f_{cr} already forms a low-variance manifold, deviations from this manifold can be efficiently simulated through stochastic perturbations. Given f_{cr}, we draw a noise vector ϵ from an i.i.d. Gaussian distribution $\mathcal{N}(0, \sigma^2)$ and synthesize a perturbed representation through additive modulation:

$$f_p = f_{cr} + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{7}$$

This formulation provides two advantages. First, perturbations in a compact frequency-conditioned space amplify subtle discrepancies between real and shifted features, producing pseudo-fake samples that mimic the off-manifold patterns commonly induced by generative models. Second, the use of isotropic noise avoids introducing model-specific artifacts, ensuring that the resulting negative samples remain paradigm-agnostic. These perturbed features are used solely during training to guide the discriminator toward learning a stable and generalizable decision boundary between real and non-real regions of the embedding space.

Discriminator

The discriminator D_χ is responsible for learning a decision boundary that separates real compact representations from the perturbed pseudo-negative samples. During training, both f_{cr} and its perturbed counterpart f_p are fed into D_χ, which outputs positive scores for real features and negative scores for pseudo-fake features. Since all features lie in the same low-dimensional space, the discriminator focuses solely on deviations from the real-feature manifold rather than on semantic or generator-specific cues.

Before classification, an adaptor maps the input features to a unified embedding dimension to stabilize optimization. We use a lightweight fully connected layer, which we found to be sufficient for preserving discriminative structure while avoiding unnecessary model capacity. A two-layer MLP then serves as the classification head, producing the normality score $D_\chi(f_{cr,p})$. Real and pseudo-negative samples are processed in parallel but independently, without feature-level coupling, ensuring that the decision boundary is shaped purely by the geometry of the compact representation space. Empirically, mapping features to 64 dimensions yields the best trade-off between stability and generalization.

Experiments

In this section, we conduct extensive experiments to evaluate RealNet across diverse generative paradigms, compare it against state-of-the-art detectors, assess its robustness under domain shifts, analyze its efficiency, and perform detailed ablations.

Experimental Settings

We outline the datasets, baseline configurations, evaluation metrics, and implementation details used to assess RealNet.

Datasets We evaluate RealNet across a broad spectrum of generative paradigms, including GAN, diffusion, and VAR-based models. Concretely, we consider ProGAN (Karras et al. 2018), BigGAN (Brock, Donahue, and Simonyan 2019), CycleGAN (Zhu et al. 2017), StarGAN (Choi et al. 2018), ADM (Dhariwal and Nichol 2021), Glide (Nichol et al. 2022), Midjourney (Midjourney 2023), SD v1.4/v1.5 (Rombach et al. 2022), VQDM (Gu et al. 2022), Wukong (wukong 2023), DALLE2 (Ramesh et al. 2022), VAR (Tian et al. 2024), CAR (Yao et al. 2024), CoDe (Chen et al. 2024b), FlowAR (Ren et al. 2024), and Infinity (Han et al. 2024). For GAN and diffusion models, we use forged images from standard public datasets (Wang et al. 2020). Because recent VAR-based forgeries are not publicly available, we additionally construct a dataset using their official pre-trained models. To evaluate robustness under distribution shifts, we further assemble a medical forgery dataset containing AI-generated and real radiology and pathology images, representing high-risk real-world scenarios.

Baselines We compare RealNet with a comprehensive set of state-of-the-art forgery detectors, including CNNSpot (Wang et al. 2020), FreDect (Frank et al. 2020), GramNet (Liu, Qi, and Torr 2020), LGrad (Tan et al. 2023), UniFD (Ojha, Li, and Lee 2023), NPR (Tan et al. 2024), Fatformer (Liu et al. 2024), AEROBLADE (Ricker, Lukovnikov, and Fischer 2024), DRCT (Chen et al. 2024a), D^3 (Yang et al. 2025), and FIRE (Chu et al. 2025). Following the protocol in (Wang et al. 2020; Ojha, Li, and Lee 2023), we either adopt publicly released checkpoints (Zhong et al. 2023; Tan et al. 2024; Liu et al. 2024) for the baselines or train them on ProGAN-generated images using official implementations (Yang et al. 2025; Chu et al. 2025). DRCT is evaluated with its diffusion-trained official checkpoint, which is a constraint inherent to its model design. RealNet is trained solely on real images by design, and for fairness, the amount of real data is matched to that used by the baselines.

Evaluation Metrics Following previous work (Wang et al. 2020; Ojha, Li, and Lee 2023), we report accuracy (Acc) and average precision (AP) for all methods. The decision threshold for Acc is calibrated using the protocol in (Wang et al. 2020; Ojha, Li, and Lee 2023). To summarize performance over all generative models, we also compute the mean accuracy (mAcc) and mean average precision (mAP).

Implementation Details Training proceeds in two stages: pre-training the Real Pattern Extractor (RPE) and subse-

Methods	Ref	GAN				Diffusion								VAR					mAP
		Pro-GAN	Big-GAN	Cycle-GAN	Star-GAN	ADM	Glide	Mid-journey	SD v1.4	SD v1.5	VQDM	Wu-kong	DALL E2	VAR	CAR	CoDe	Flow-AR	Infinity	
CNNSpot	CVPR2020	100.0	84.51	93.48	98.15	71.10	66.18	55.93	56.91	57.31	61.96	52.90	50.53	42.93	52.53	45.67	53.92	38.23	63.66
FreDect	ICML2020	99.99	93.62	84.77	99.49	61.76	52.91	46.08	37.83	37.76	85.10	39.58	38.20	68.45	61.37	76.77	70.18	39.27	64.30
GramNet	CVPR2020	100.0	62.34	74.82	100.0	57.07	55.18	58.78	63.11	63.31	54.19	61.01	53.74	45.58	42.81	79.99	77.91	79.58	66.44
LGrad	CVPR2023	100.0	89.10	93.78	99.98	67.01	82.42	73.57	63.32	63.73	71.88	61.26	84.13	58.96	59.58	77.60	89.27	72.19	76.93
UniFD	CVPR2023	100.0	99.27	99.80	99.37	89.80	88.04	49.72	68.63	68.07	97.53	78.44	66.06	68.95	93.31	82.60	90.79	39.90	81.19
NPR	CVPR2024	99.95	84.40	97.83	100.0	79.14	86.55	83.84	84.37	84.38	80.84	77.63	79.56	41.39	40.74	91.04	89.75	65.67	80.42
Fatformer	CVPR2024	100.0	99.98	100.0	100.0	91.73	95.99	62.76	81.12	81.09	96.99	85.86	81.84	66.37	86.80	99.26	99.25	63.94	87.82
AEROBLAD	CVPR2024	46.48	42.14	40.87	43.38	87.42	97.96	99.84	98.68	98.87	78.42	99.07	98.69	35.69	36.86	41.38	40.87	45.59	66.60
DRCT	ICML2024	91.03	93.51	98.68	96.29	88.96	94.64	97.03	99.65	99.49	96.54	99.37	97.67	60.20	69.92	74.08	82.91	92.32	90.13
D^3	CVPR2025	100.00	98.18	99.91	98.84	95.86	92.42	80.15	84.83	84.96	93.38	85.44	81.34	87.65	86.92	99.21	84.18	84.60	<u>90.46</u>
FIRE	CVPR2025	100.00	97.69	98.29	99.03	93.62	90.13	82.76	86.42	86.02	90.76	86.05	82.79	84.10	84.97	99.34	83.95	86.70	90.15
Ours	-	77.81	96.46	99.78	100.0	97.96	95.00	90.31	92.55	93.05	97.99	92.06	97.90	95.28	95.17	99.97	91.87	91.46	**94.39**

Table 1: **Cross-model average precision (%) comparison with state-of-the-art detectors.** The best and second-best results are marked in **bold** and <u>underline</u>, respectively.

Methods	Ref	GAN				Diffusion								VAR					mAcc
		Pro-GAN	Big-GAN	Cycle-GAN	Star-GAN	ADM	Glide	Mid-journey	SD v1.4	SD v1.5	VQDM	Wu-kong	DALL E2	VAR	CAR	CoDe	Flow-AR	Infinity	
CNNSpot	CVPR2020	99.99	81.13	86.34	92.75	67.31	66.30	54.44	58.58	58.67	62.37	54.83	52.50	50.63	55.25	50.85	52.58	50.13	64.39
FreDect	ICML2020	99.75	86.78	81.42	97.57	68.33	63.44	50.18	50.04	50.01	78.28	50.03	50.00	72.13	68.33	69.53	62.50	50.53	67.58
GramNet	CVPR2020	100.0	68.05	75.28	100.0	60.41	58.42	62.50	69.78	69.98	58.37	66.06	56.55	53.80	48.23	70.73	70.33	70.70	68.19
LGrad	CVPR2023	99.80	82.40	85.92	99.30	61.95	73.07	67.82	64.65	65.04	69.27	60.77	76.25	56.90	56.38	73.55	72.13	73.50	72.86
UniFD	CVPR2023	99.86	95.78	98.33	96.65	80.78	79.98	50.42	66.43	66.33	91.53	72.67	64.95	63.75	84.60	74.10	80.78	50.03	77.47
NPR	CVPR2024	99.86	85.05	96.40	99.82	75.58	88.61	83.26	83.93	84.32	79.53	75.13	86.60	50.00	50.00	93.90	91.83	70.40	82.01
Fatformer	CVPR2024	99.96	99.63	99.92	100.0	83.48	89.65	58.41	71.75	72.05	90.43	76.98	73.95	60.48	77.65	96.30	95.98	60.51	82.77
AEROBLAD	CVPR2024	50.04	50.10	50.08	53.03	80.32	93.02	98.84	97.83	97.91	73.88	98.14	95.95	50.03	50.03	52.70	51.20	60.16	70.78
DRCT	ICML2024	83.31	86.45	94.51	90.30	80.77	89.33	91.63	97.29	96.73	90.18	96.43	93.35	58.35	65.15	67.63	73.88	83.79	84.65
D^3	CVPR2025	99.89	97.84	98.35	97.60	77.36	74.21	70.44	73.73	73.82	89.72	72.32	70.07	80.51	81.90	98.92	80.83	82.65	83.54
FIRE	CVPR2025	96.93	83.36	86.13	95.08	85.66	84.42	77.88	85.24	85.18	79.81	84.78	83.54	79.49	79.38	99.12	79.45	80.02	<u>85.03</u>
Ours	-	69.75	90.33	97.99	99.90	95.47	89.68	82.39	87.33	87.75	94.60	86.72	95.00	88.73	88.38	99.53	84.13	84.56	**89.54**

Table 2: **Cross-model accuracy (%) comparison with state-of-the-art detectors.** Notations follow those in Table 1.

quently training RealNet. RPE is trained as an image denoiser using real clean–noisy pairs $(\boldsymbol{x}, \boldsymbol{y})$. Following (Arjovsky, Chintala, and Bottou 2017; LI et al. 2017), the restorer R, generator G, and discriminator D are jointly optimized under the WGAN-GP framework to ensure stable adversarial training. For each update of R and G, the discriminator is updated three times. The weights of R and G are initialized with He initialization (He et al. 2015), while D follows the initialization strategy in (Yu et al. 2017) using a normal distribution with standard deviation 0.02. Adam is used for all three modules, with momentum terms set to (0.9, 0.999), (0.5, 0.9), and (0.5, 0.9), and learning rates of 1×10^{-4}, 1×10^{-4}, and 2×10^{-4}, respectively. Learning rates are reduced by 50% every 10 epochs. We set $\lambda_1 = 1000$ and $\lambda_2 = 10$, and adopt the same $\mathcal{L}_{GAN}$ hyper-parameters as in (LI et al. 2017).

For RealNet training, we use the real-image subset of the training split in (Wang et al. 2020). SARR features are resized to 256×256 before entering the Feature Transformer, and the resulting compact real representation has a dimensionality of 64. The Feature Perturbator adds Gaussian noise with distribution $\varepsilon \sim \mathcal{N}(0, 0.008)$, and the adaptor in the discriminator is implemented as a single linear layer. We train RealNet for 100 epochs with a batch size of 8 using Adam, an initial learning rate of 1×10^{-4}, and cosine annealing scheduling. All experiments are implemented in PyTorch and conducted on a single NVIDIA GeForce RTX 4090 GPU.

Comparison with the State of the Arts

Here we compare RealNet with state-of-the-art detectors in terms of detection performance, cross-paradigm generalization, computational efficiency, and robustness under domain shifts.

Quantitative Results Across all test sets (Tables 1 and 2), RealNet achieves the highest accuracy (Acc) and average precision (AP) among all competing detectors, with improvements of 4.51% in mean accuracy (mAcc) and 3.93% in mean average precision (mAP). These gains demonstrate the effectiveness of learning a forgery-aware representation entirely from real images, which reduces reliance on synthetic supervision and mitigates overfitting to semantic content or generator-specific cues.

RealNet also exhibits strong cross-paradigm generalization. UniFD, trained on ProGAN-generated forgeries, performs well on GAN data but drops sharply on diffusion and VAR models. DRCT, optimized on Stable Diffusion outputs, exhibits the opposite pattern, strong diffusion performance but limited transfer to GAN and VAR domains. AEROBLADE, although training-free, depends on autoencoder reconstruction residuals and therefore aligns best with latent diffusion architectures. In contrast, RealNet maintains consistently high performance across paradigms due to its compact and paradigm-agnostic representation. On challenging

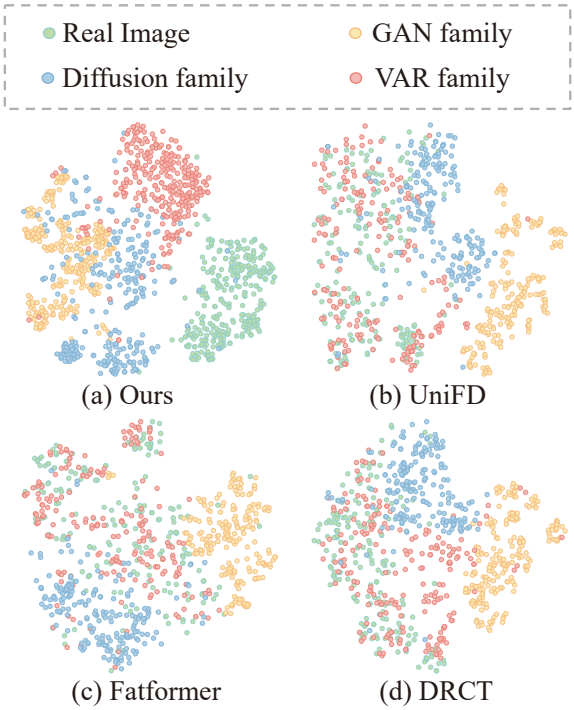

Figure 3: **Embedding space visualization.** t-SNE plots of RealNet and baseline detectors. Baselines show dispersed and generator-biased clusters; RealNet produces compact real embeddings and clearly separated fake regions across generative paradigms.

VAR-based generators, RealNet surpasses supervised detectors by more than 6.24% mAP, demonstrating its strong generalization to diverse and previously unseen generative mechanisms.

Qualitative Results Figure 3 presents t-SNE (Van der Maaten and Hinton 2008) visualizations of feature distributions from RealNet and baseline detectors. Existing methods produce dispersed real-image features and separate fake clusters only when the forgery type resembles their training data, revealing strong generator-specific biases and limited transfer to unseen architectures. RealNet yields a markedly different geometry: real samples form a compact cluster, while fakes from GAN, diffusion, and VAR models occupy distinct, well-separated regions. This confirms the effectiveness of our semantic-agnostic representation learning and noise-based pseudo-negative generation, which together enable stable decision boundaries and strong generalization to previously unseen generative paradigms.

Efficiency Test We assess the computational efficiency of RealNet, an important factor for real-world deployment. Table 3 reports inference throughput on two widely used GPU platforms (RTX 4090 and 2080Ti). RealNet achieves the highest processing speed among all compared methods, benefiting from its compact architecture and the absence of heavy reconstruction modules or large pre-trained backbones. These results indicate that RealNet is well suited for real-time and resource-limited settings, complementing its strong generalization performance.

Model	FPS (4090)	FPS (2080Ti)	mAcc	mAP
NPR	18.4	8.1	82.01	80.42
Fatformer	9.6	5.2	82.77	87.82
DRCT	11.8	4.8	84.65	90.13
D^3	10.4	5.5	83.54	90.46
FIRE	7.1	4.4	85.03	90.15
Ours	**24.5**	**10.1**	**89.54**	**94.39**

Table 3: **Inference efficiency comparison.** We report throughput (FPS) and overall performance (mAcc/mAP, %) on two GPU platforms.

Model	mAcc	mAP
NPR	72.03	76.14
Fatformer	73.91	76.72
DRCT	76.34	79.48
D^3	77.60	80.57
FIRE	75.12	79.17
Ours	**81.20**	**84.74**

Table 4: **Domain transfer evaluation on the medical forgery dataset.** Results are reported in accuracy and average precision (%).

Domain Transfer Evaluation We further examine RealNet's robustness to domain shifts on a medical forgery dataset containing real and AI-generated radiology and pathology images. These images exhibit characteristics that differ markedly from natural-image distributions and are associated with high-stakes risks such as diagnostic errors and clinical misuse (Albahli and Nawaz 2024; Alsaheel et al. 2023). As shown in Table 4, RealNet achieves a clear performance advantage over prior detectors. Existing methods degrade substantially because their feature representations entangle semantic cues or rely on generator-specific artifacts that do not transfer to the medical domain. In contrast, RealNet's real-only training paradigm and disentangled representation space yield strong domain generalization without requiring medical synthetic data during training.

Ablation Analysis

We conduct comprehensive ablations to evaluate the contribution of each component in RealNet, with results summarized in Table 5. Replacing the Real Pattern Extractor (RPE) with other restoration or feature-extraction models (Tu et al. 2022; Chen et al. 2022a; He et al. 2016) leads to substantial performance degradation (e.g., NAFNet causes a 12.25%/13.49% drop in mAcc/mAP), indicating that these alternatives fail to produce compact and perturbation-sensitive real representations, whereas RPE's dual adversarial design is crucial for learning such properties. Removing the Feature Transformer (FT) or substituting it with simpler mappings (a 1×1 convolution or a single bilinear layer) likewise reduces accuracy, confirming that FT's frequency-domain transformation and structured compression are important for forming a discriminative compact

Component	Settings	mAcc	mAP
Transformer	w/o trans.	58.15	56.23
	vanilla trans.	60.33	55.76
	feature compression	**89.54**	**94.39**
Extractor	ResNet 50	55.03	52.22
	MAXIM	83.94	87.43
	NAFNet	77.29	80.90
	noise sensitive restoration	**89.54**	**94.39**
Variance	0.004	88.40	93.26
	0.006	88.51	92.61
	0.008	**89.54**	**94.39**
	0.010	86.95	90.87
Discriminator	1 layer, 32 dim	83.24	88.52
	1 layer, 64 dim	**89.54**	**94.39**
	1 layer, 128 dim	88.17	93.18
	2 layers, 64 dim	89.10	93.67

Table 5: **Ablation study of the main components.** Results reflect mean accuracy and average precision (%) across all test sets.

space and enabling effective pseudo-negative synthesis. We further vary the noise variance in the Perturbator and observe that a variance of 0.008 yields the most reliable performance, while smaller or larger perturbations weaken the decision boundary. Finally, experiments on the discriminator adaptor show that a single-layer MLP with a 64-dimensional output achieves the best results, suggesting that lightweight adaptation before classification is sufficient and beneficial for stable optimization.

Conclusion

In this paper, we presented RealNet, an unsupervised framework that learns a semantic-agnostic and model-agnostic representation space directly from real images, without relying on synthetic forgeries. By disentangling forgery-irrelevant semantics and suppressing model-specific artifacts, RealNet forms a compact and perturbation-sensitive feature space that supports stable and generalizable detection. Extensive experiments show that RealNet consistently surpasses state-of-the-art detectors across GAN, diffusion, and emerging VAR paradigms, and remains robust under severe distribution shifts such as medical image forgery, while requiring substantially lower computational cost. These results highlight real-only representation learning as a practical and scalable direction for reliable AI-generated content detection in high-impact real-world settings.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62176010), the Beijing Natural Science Foundation (No. 4252029), the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No. VRLAB 2024B02), the National Natural Science Foundation of China (No. 62572059), and the Fundamental Research Funds for the Central Universities (No. 2253200009).

References

Abdelhamed, A.; Lin, S.; and Brown, M. S. 2018. A High-Quality Denoising Dataset for Smartphone Cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Albahli, S.; and Nawaz, M. 2024. MedNet: Medical deepfakes detection using an improved deep learning approach. *Multimedia Tools and Applications*, 83(16): 48357–48375.

Alsaheel, A.; Alhassoun, R.; Alrashed, R.; Almatrafi, N.; Almallouhi, N.; and Albahli, S. 2023. Deep Fakes in Healthcare: How Deep Learning Can Help to Detect Forgeries. *Computers, Materials & Continua*, 76(2).

Anaya, J.; and Barbu, A. 2018. RENOIR-A benchmark dataset for real noise reduction evaluation. *Journal of Visual Communication and Image Representation*, 144–154.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*, 214–223. PMLR.

Bi, X.; Liu, B.; Yang, F.; Xiao, B.; Li, W.; Huang, G.; and Cosman, P. C. 2023. Detecting Generated Images by Real Images Only. arXiv:2311.00962.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*.

Chen, B.; Zeng, J.; Yang, J.; and Yang, R. 2024a. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*.

Chen, L.; Chu, X.; Zhang, X.; and Sun, J. 2022a. Simple Baselines for Image Restoration. In *Computer Vision – ECCV 2022*, 17–33. Cham: Springer Nature Switzerland.

Chen, Y.; Mancini, M.; Zhu, X.; and Akata, Z. 2022b. Semi-supervised and unsupervised deep visual learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(3): 1327–1347.

Chen, Z.; Ma, X.; Fang, G.; and Wang, X. 2024b. Collaborative Decoding Makes Visual Auto-Regressive Modeling Efficient. *arXiv preprint arXiv:2411.17787*.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chu, B.; Xu, X.; Wang, X.; Zhang, Y.; You, W.; and Zhou, L. 2025. Fire: Robust detection of diffusion-generated images via frequency-guided reconstruction error. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12830–12839.

Dao, Q.; Phung, H.; Nguyen, B.; and Tran, A. 2023. Flow Matching in Latent Space. *arXiv e-prints*, arXiv:2307.08698.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, 8780–8794.

Durall, R.; Keuper, M.; and Keuper, J. 2020. 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 (CVPR)*.

Dzanic, T.; Shah, K.; and Witherden, F. 2020. Fourier Spectrum Discrepancies in Deep Network Generated Images. In *Advances in Neural Information Processing Systems*, volume 33, 3022–3032.

Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; and Holz, T. 2020. Leveraging Frequency Analysis for Deep Fake Image Recognition. In *Proceedings of the 37th International Conference on Machine Learning*, 3247–3258.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022. Vector Quantized Diffusion Model for Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10696–10706.

Han, J.; Liu, J.; Jiang, Y.; Yan, B.; Zhang, Y.; Yuan, Z.; Peng, B.; and Liu, X. 2024. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*.

Hao, P.; Li, S.; Wang, H.; Kou, Z.; Zhang, J.; Yang, G.; and Zhu, L. 2025. Surgery-R1: Advancing Surgical-VQLA with Reasoning Multimodal Large Language Model via Reinforcement Learning. *arXiv preprint arXiv:2506.19469*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, Z.; Chen, P.-Y.; and Ho, T.-Y. 2024. Rigid: A training-free and model-agnostic framework for robust ai-generated image detection. *arXiv preprint arXiv:2405.20112*.

Jeong, Y.; Kim, D.; Ro, Y.; Kim, P.; and Choi, J. 2022. Fingerprintnet: Synthesized fingerprints for generated image detection. In *European Conference on Computer Vision*, 76–94. Springer.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *Proceedings of International Conference on Learning Representations (ICLR) 2018*.

Kingma, D. P.; and Dhariwal, P. 2018. Glow: Generative Flow with Invertible 1x1 Convolutions. In *Advances in Neural Information Processing Systems*, volume 31.

LI, C.; Xu, T.; Zhu, J.; and Zhang, B. 2017. Triple Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Li, S.; Ma, W.; Guo, J.; Xu, S.; Li, B.; and Zhang, X. 2024. Unionformer: Unified-learning transformer with multi-view representation for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12523–12533.

Li, S.; Xing, Z.; Wang, H.; Hao, P.; Li, X.; Liu, Z.; and Zhu, L. 2025. Toward Medical Deepfake Detection: A Comprehensive Dataset and Novel Method. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 626–637. Springer.

Li, S.; Xu, S.; Ma, W.; and Zong, Q. 2021. Image manipulation localization using attentional cross-domain CNN features. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9): 5614–5628.

Liu, H.; Tan, Z.; Tan, C.; Wei, Y.; Wang, J.; and Zhao, Y. 2024. Forgery-aware Adaptive Transformer for Generalizable Synthetic Image Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, Y.; Qin, Z.; Anwar, S.; Ji, P.; Kim, D.; Caldwell, S.; and Gedeon, T. 2021. Invertible Denoising Network: A Light Solution for Real Noise Removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13365–13374.

Liu, Z.; Qi, X.; and Torr, P. H. 2020. Global Texture Enhancement for Fake Face Detection in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Marra, F.; Gragnaniello, D.; Verdoliva, L.; and Poggi, G. 2019. Do gans leave artificial fingerprints? In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, 506–511. IEEE.

Midjourney. 2023. https://www.midjourney.com/home/.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of the 39th International Conference on Machine Learning*, 16784–16804.

Ojha, U.; Li, Y.; and Lee, Y. J. 2023. Towards Universal Fake Image Detectors That Generalize Across Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24480–24489.

Prezja, F.; Paloneva, J.; Pölönen, I.; Niinimäki, E.; and Äyrämö, S. 2022. DeepFake knee osteoarthritis X-rays from generative adversarial neural networks deceive medical experts and offer augmentation potential to automatic classification. *Scientific Reports*, 12(1): 18573.

Qu, C.; Liu, C.; Liu, Y.; Chen, X.; Peng, D.; Guo, F.; and Jin, L. 2023. Towards robust tampered text detection in document image: New dataset and new solution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5937–5946.

Qu, C.; Zhong, Y.; Guo, F.; and Jin, L. 2024a. Omni-IML: Towards Unified Image Manipulation Localization. *arXiv preprint arXiv:2411.14823*.

Qu, C.; Zhong, Y.; Guo, F.; and Jin, L. 2025. Revisiting tampered scene text detection in the era of generative AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 694–702.

Qu, C.; Zhong, Y.; Liu, C.; Xu, G.; Peng, D.; Guo, F.; and Jin, L. 2024b. Towards modern image manipulation localization: A large-scale dataset and novel methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10781–10790.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *ArXiv*, abs/2204.06125.

Razavi, A.; van den Oord, A.; and Vinyals, O. 2019. Generating Diverse High-Fidelity Images with VQ-VAE-2. In *Advances in Neural Information Processing Systems*.

Ren, S.; Yu, Q.; He, J.; Shen, X.; Yuille, A.; and Chen, L.-C. 2024. FlowAR: Scale-wise Autoregressive Image Generation Meets Flow Matching. *arXiv preprint arXiv:2412.15205*.

Ricker, J.; Lukovnikov, D.; and Fischer, A. 2024. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9130–9140.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 234–241. Cham: Springer International Publishing.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024. 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*, 28130–28139.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; and Wei, Y. 2023. Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12105–12114.

Tian, K.; Jiang, Y.; Yuan, Z.; Peng, B.; and Wang, L. 2024. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. *arXiv e-prints*, arXiv:2404.02905.

Tian, K.; Jiang, Y.; Yuan, Z.; Peng, B.; and Wang, L. 2024. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. arXiv:2404.02905.

Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; and Li, Y. 2022. MAXIM: Multi-Axis MLP for Image Processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5769–5780.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; Hu, H.; Chen, H.; and Li, H. 2023. DIRE for Diffusion-Generated Image Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 22445–22455.

wukong. 2023. https://xihe.mindspore.cn/modelzoo/wukong.

Xu, J.; Li, H.; Liang, Z.; Zhang, D.; and Zhang, L. 2018. Real-world noisy image denoising: A new benchmark. *arXiv preprint arXiv:1804.02603*.

Yang, Y.; Qian, Z.; Zhu, Y.; Russakovsky, O.; and Wu, Y. 2025. Dˆ 3: Scaling Up Deepfake Detection by Learning from Discrepancy. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 23850–23859.

Yao, Z.; Li, J.; Zhou, Y.; Liu, Y.; Jiang, X.; Wang, C.; Zheng, F.; Zou, Y.; and Li, L. 2024. Car: Controllable autoregressive modeling for visual generation. *arXiv preprint arXiv:2410.04671*.

Yu, N.; Davis, L. S.; and Fritz, M. 2019. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Yu, Y.; Gong, Z.; Zhong, P.; and Shan, J. 2017. Unsupervised representation learning with deep convolutional neural network for remote sensing images. In *Image and Graphics: 9th International Conference, ICIG 2017, Shanghai, China, September 13-15, 2017, Revised Selected Papers, Part II 9*, 97–108. Springer.

Yue, Z.; Zhao, Q.; Zhang, L.; and Meng, D. 2020. Dual Adversarial Network: Toward Real-world Noise Removal and Noise Generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zhang, L.; Li, S.; Ma, W.; and Zha, H. 2025. TrueMoE: Dual-Routing Mixture of Discriminative Experts for Synthetic Image Detection. *arXiv preprint arXiv:2509.15741*.

Zhang, X.; Karaman, S.; and Chang, S.-F. 2019a. Detecting and Simulating Artifacts in GAN Fake Images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–6.

Zhang, X.; Karaman, S.; and Chang, S.-F. 2019b. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, 1–6. IEEE.

Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; and Yu, N. 2021. Multi-Attentional Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2185–2194.

Zhong, N.; Xu, Y.; Li, S.; Qian, Z.; and Zhang, X. 2023. PatchCraft: Exploring Texture Patch for Efficient AI-generated Image Detection. *arXiv e-prints*, arXiv:2311.12397.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.