Decision-Driven Orthogonal Learning with Complementary Feature Mining for Robust Synthetic Image Detection

Kai Li[1,4], Wei Wang[2], Linchao Zhang[3], Siying Zhu[3], Wenqi Ren[2,4*],

[1]School of Computer Science and Engineering, Sun Yat-sen University
[2]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University
[3]Information Science Academy, China Electronics Technology Group Corporation
[4]The State KeyLaboratory of Blockchain and Data Security, Zhejiang University
likai63@mail2.sysu.edu.cn, {renwq3, wangwei29}@mail.sysu.edu.cn, zsiyingstu@gmail.com, hune213@163.com

Abstract

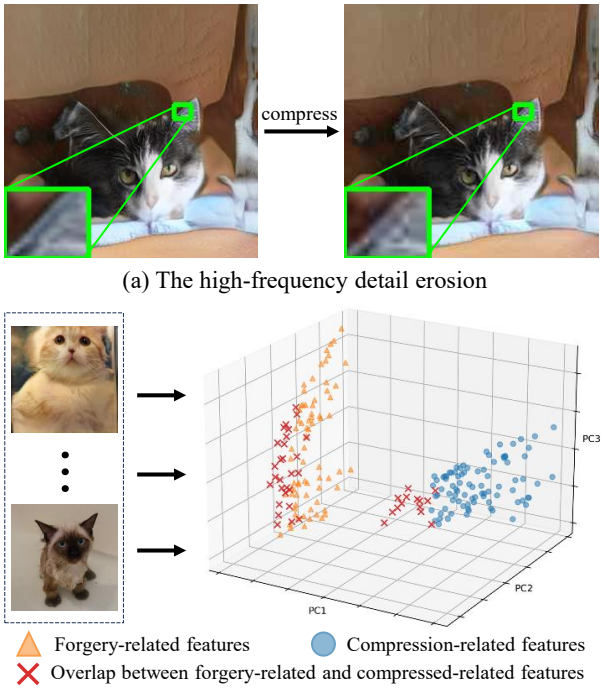
The widespread and inconsistent compression applied by Online Social Networks severely degrades the performance of synthetic image detectors. We attribute this degradation to two main issues: 1) the model confuses forgery artifacts with compression artifacts, and 2) compression erodes crucial discriminative high-frequency details. Existing methods suppress compression features during training but overlook the overlap between compression features and forgery-related features, leading to the unintended removal of forgery traces. To address artifact confusion, we introduce a Decision-Driven Orthogonal Constraint, which defines a classification decision axis pointing from the real class centroid to the forged class centroid. This constraint enforces compression artifacts to be orthogonal to the decision axis, mitigating their interference with forgery detection without entirely removing them, thus preventing the suppression of forgery-related features. To mitigate the erosion of high-frequency details, we propose to mine complementary forgery cues from both low-frequency information and compressed high-frequency components. A bidirectional update strategy and an adaptive global-local modulator are proposed to facilitate the utilization of forgery cues. Extensive experiments demonstrate that our method achieves state-of-the-art generalization performance in challenging open-world detection scenarios.

Introduction

The rapid evolution of generative models (Goodfellow et al. 2014; Ho, Jain, and Abbeel 2020) has resulted in an increase of synthetic images on Online Social Networks (OSNs), raising serious concerns regarding misinformation and fraudulent content. Despite significant advancements (Tan et al. 2024a; Liu et al. 2024; Tan et al. 2024b) in synthetic image detection, performance degrades when the detector is deployed in real-world scenarios. This degradation is primarily caused by the inconsistent image compression employed by OSNs (Sun et al. 2016).

Deploying detectors in this scenario exposes them to two fundamental and significant challenges. First, the overlap of forgery and compression artifacts. Models are simultaneously exposed to forgery artifacts and compression artifacts, where these artifacts may overlap in the feature space. Consequently, the model often confuses compression artifacts with forgery artifacts, leading to a significant drop in detection accuracy. Second, the erosion of high-frequency details. Compression algorithms inherently discard high-frequency information to reduce file size, which contains the most discriminative forgery traces (Tan et al. 2024b). The removal of these critical high-frequency traces leads to a dramatic performance drop for detectors.

Existing methods addressing these challenges can be broadly summarized into two main categories. The first cat-



(a) The high-frequency detail erosion

(b) The information overlap

Figure 1: The impact of compression. (a) Compression perturbations degrade image details, especially in the high-frequency components. (b) Gradient reversal-based detectors may remove critical forgery-related features when eliminating compression-related information.

*Corresponding authors: Wenqi Ren.

egory (Le and Woo 2023; Wu et al. 2022; Le and Woo 2022) relies on paired data (original and its compressed versions) to train compression-insensitive models. However, these approaches are inherently limited by the lack of such paired data in real-world scenarios, which limits the practical application of the model. An alternative (Tao et al. 2025) employs the gradient reversal strategy (Ganin and Lempitsky 2015) to learn features that are robust to compression, requiring only a small number of paired samples. While this approach is effective in suppressing compression-related artifacts, it results in the removal of informative features. Since forgery and compression artifacts may share similar feature representations, this process can unintentionally eliminate cues that are critical for forgery detection, as shown in Figure 1. Moreover, the erosion of discriminative high-frequency information by compression is overlooked.

To address these two issues, we propose a framework based on two complementary principles. First, to resolve the confusion issue of forgery and compression features, we introduce a Decision-Driven Orthogonal Constraint. A straightforward strategy is to enable the compression perturbation and the forgery artifact to be orthogonal. The compression perturbations can be obtained by subtracting features of the original and compressed images. However, the corresponding forgery artifact is inaccessible using the same manner. The reason is that such a computation necessitates access to ground-truth pairs of real images and their synthetic counterparts, which is generally unavailable. Given this limitation, we propose to operate directly on the decision-making process, rather than attempting to model the underlying artifacts. Specifically, our method explicitly defines a classification decision axis, which points from the real class centroid to the forged class centroid. Subsequently, we enforce compression-related features to be orthogonal to this decision axis. This forces the compression perturbation to have zero projection onto the decision axis, thus preventing the compression perturbation from affecting the classification decision of the detector. This approach enables the detector to achieve compression robustness without removing features, thus preserving forgery-relevant information.

Second, to mitigate the erosion of high-frequency details, our method is designed to exploit forgery cues from two complementary sources: the robust low-frequency structures of an image and the information in the compressed high-frequency components. We introduce a hybrid framework that combines the complementary strengths of different architectures. Specifically, it leverages the ability of Vision Transformers (Dosovitskiy et al. 2021) to model global, low-frequency patterns (Park and Kim 2022; Si et al. 2022) and that of CNNs to capture local, high-frequency details (Wang et al. 2020a; Si et al. 2022). A core component of this framework is a bidirectional update strategy that enables information interaction between the low- and high-frequency processing streams. This facilitates the fusion of artifacts from both frequencies, and an adaptive global-local modulator is proposed to further refine the fused features.

The main contributions of our proposed method can be summarized as follows:

- We propose a Decision-Driven Orthogonal Constraint that enhances compression robustness by enforcing the compression perturbation to be orthogonal to the introduced decision axis, without removing the compression perturbation. This approach preserves forgery-related features that might be lost if perturbations are removed.
- We propose to exploit forgery cues from two complementary sources: the robust low-frequency components and the information in the compressed high-frequency components. To effectively integrate information from these frequency domains, we introduce a bidirectional update strategy and an adaptive global-local modulator.
- Extensive experiments on multiple benchmarks demonstrate that our method surpasses state-of-the-art detectors in various open-world scenarios, demonstrating its robustness and practical value in open-world scenarios.

Related Work

A significant amount of research has been dedicated to improving the generalization performance of detection models. Early methods mainly focused on visually salient artifacts introduced during image synthesis (Li et al. 2020; Kaede and Toshihiko 2022). However, as generative models—especially Generative Adversarial Networks (Goodfellow et al. 2014) and diffusion models (Ho, Jain, and Abbeel 2020)—have rapidly advanced, these artifacts have become increasingly subtle and difficult to perceive. To adapt to the evolving generation techniques, researchers have proposed various advanced detection strategies, including frequency-domain analysis, semantic feature extraction, and reconstruction error-based methods. Frequency-domain-based methods (Frank et al. 2020; Durall, Keuper, and Keuper 2020; Tan et al. 2024b; Li et al. 2025) focus on detecting abnormalities in the frequency spectrum of images, where forgery clues may be more evident than in the spatial domain. For example, some studies (Frank et al. 2020; Durall, Keuper, and Keuper 2020) have shown that upsampling operations used in generators can leave unique grid-like artifacts. NPR (Tan et al. 2024a) rethinks these upsampling-induced artifacts from a spatial perspective, introducing the concept of neighboring pixel relationships.

Semantic feature-based approaches (Ojha, Li, and Lee 2023; Liu et al. 2024; Tan et al. 2023; Shi et al. 2025) aim to construct a generalized feature space for detection. For instance, UnivFD (Ojha, Li, and Lee 2023) builds a universal linear classifier using a pre-trained CLIP-ViT (Radford et al. 2021), significantly improving cross-model generalization. FatFormer (Liu et al. 2024) further enhances this framework by incorporating forgery-aware adapters to better capture localized forged traces. LGrad (Tan et al. 2023) utilizes gradient feature maps to capture artifacts. Reconstruction-based methods exploit the difference between the input image and its reconstructed version as an auxiliary signal. For example, RECCE (Cao et al. 2022) reconstructs only real images, causing the reconstructor to fail when handling forged ones. DIRE (Wang et al. 2023) observes that real images can be better reconstructed by diffusion models than fake ones, and thus introduces a pre-trained diffusion model as the reconstructor. LaRE2 (Luo et al. 2024) further simplifies this ap-

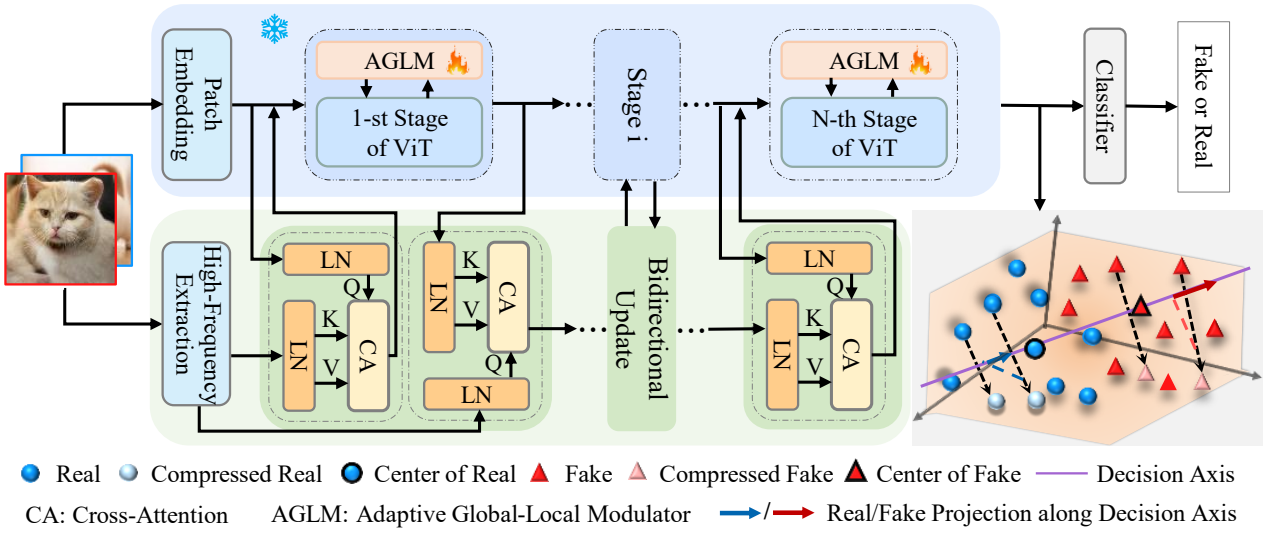


Figure 2: The overview of our proposed method. The input image is processed by both a ViT-based backbone for extracting low-frequency features and a convolutional branch dedicated to extracting high-frequency information. A bidirectional update strategy is employed to fuse information from the two branches. Only low-frequency features are updated in the last fusion. The fused features are further refined through an adaptive global-local modulator. These refined features are then subject to a decision-driven orthogonality constraint, enforcing the detector to learn compression-robust features.

proach, showing that even single-step reconstructions can effectively distinguish real from generated images without requiring the full diffusion process used in DIRE.

However, a major challenge faced by these methods is that compression on social media platforms severely degrades image details (Sun et al. 2016), especially critical high-frequency information, thereby undermining detection performance. Moreover, the compression strategies employed by OSNs are often diverse and inconsistent. Such processing fundamentally alters the data distribution of images, introducing a significant domain shift and imposing stringent robustness requirements on detectors. Recent studies have attempted to either learn compression-robust representations using large collections of original-compressed image pairs (Le and Woo 2023, 2022), or to directly suppress compression-related features (Tao et al. 2025) with limited paired data. However, the former is difficult to collect in practice, and the latter inevitably removes forgery-related features as well. To address this issue, we propose a decision-driven orthogonality constraint and a complementary low- and high-frequency interaction framework.

Method

Problem Definition

Let the initial dataset be denoted as $D = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i represents an image and $y_i \in \{0, 1\}$ denotes its corresponding label, with "0" indicating a real image and "1" a synthetic one. To construct the training set, we first divide D into two disjoint subsets: a paired subset D_p and a remaining subset D_r. Following standard protocol (Tao et al. 2025), D_p consists of a randomly sampled portion of D (20%). We then generate a compressed version of the paired subset by applying a fixed compression operator $C(\cdot)$. The compressed-paired subset D_c is defined as $D_c = \{(C(x), y) \mid (x, y) \in D_p\}$. The final training set is formed by the union $D_{train} = D_p \cup D_r \cup D_c$.

During inference, all images in the test set are compressed. The evaluation is conducted under two settings: quality-aware, where the compression quality is consistent with that of the training set, and quality-agnostic, where the compression is applied using randomly selected ratios. Our goal is to develop a detector $f(\cdot)$ that performs well on D_{train} and generalizes to real-world samples with unknown compression qualities and unseen generators.

Overview

As illustrated in Figure 2, our proposed framework adopts a dual-branch architecture designed to robustly extract forgery features from input images. The primary branch employs a pre-trained Vision Transformer (ViT) backbone to extract low-frequency features, and a dedicated convolutional branch is used to extract forgery cues (Wang et al. 2020a) within the high-frequency components. A bidirectional update strategy and a adaptive global-local modulator are introduced to fuse and refine features. After feature processing, we apply the Decision-Driven Orthogonal Constraint (DDOC) to enforce compression perturbation to be orthogonal to the classification decision axis.

Before the bidirectional update strategy, we extract high-frequency features using stacked multi-scale convolutional layers. Specifically, the input image x is processed to generate feature sequences at different scales. Each feature sequence F_h^i corresponds to a specific spatial resolution and receptive field, thus capturing multi-level detail informa-

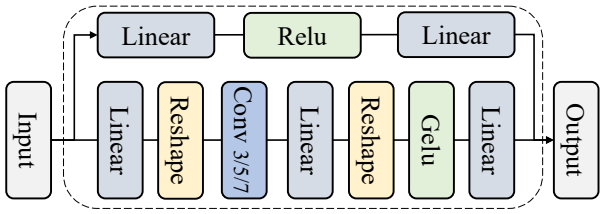


Figure 3: The adaptive global-local modulator.

tion. These multi-scale feature sequences are then concatenated to form the final high-frequency feature representation $F_h = [F_h^0; \ldots; F_h^i]$.

Bidirectional update strategy. To achieve effective fusion of high-frequency artifacts and low-frequency context, we design a bidirectional update strategy. Such a design enables the two streams to mutually guide and co-evolve during feature extraction. Specifically, it consists of two stages: high-frequency injection and high-frequency update. In the first stage, high-frequency information is injected into the ViT's main feature stream. This is achieved through a cross-attention operation, where the global features F_l from ViT serve as the Query, and the high-frequency feature set F_h acts as both the Key and Value. The output is then refined via a channel-wise attention mechanism to recalibrate feature importance. This process can be viewed as the global low-frequency features F_l actively absorbing complementary artifacts from F_h. It results in an enhanced feature representation $\hat{F}_l$. This operation can be formulated as follows:

$$\text{Attn}_{h\to l} = \text{CA}(\text{LN}(F_l), \text{LN}(F_h)), \tag{1}$$

$$\hat{F}_l = F_l + \text{ChannelAttn}(\text{Attn}_{h\to l}). \tag{2}$$

where CA denotes the cross-attention operation, and LN represents the layer normalization.

In the second stage, the high-frequency update stage is responsible for updating the high-frequency feature representation. After the low-frequency features have been enhanced through the injection process and stacked transformer blocks, the extractor performs a reverse cross-attention operation. In this stage, the high-frequency features F_h act as the Query, while the injected features $\hat{F}_l$ serve as the Key and Value. This enables F_h to focus on the discriminative high-frequency regions under the guidance of global context, thereby supplying higher-quality information for subsequent interactions. This process can be formulated as follows:

$$\text{Attn}_{l\to h} = \text{CrossAttn}(\text{LN}(F_h), \text{LN}(\hat{F}_l)), \tag{3}$$

$$\hat{F}h = F_h + \text{ChannelAttn}(\text{Attn}_{l\to h}). \tag{4}$$

By alternately performing the injection and update stages, the global low-frequency context and high-frequency information are mutually enhanced through successive interactions. After multiple rounds of such interaction, the output features effectively integrate the complementary strengths of both frequency domains.

Adaptive global-local modulator. To further refine the fused features, we introduce an adaptive global-local modulator, comprising a local refinement branch and a global processing branch. Let the input feature tensor from a multi-head attention layer in a transformer block be $F \in \mathbb{R}^{N\times C}$, where N is the sequence length and C is the feature dimension. We partition F into the class token $F_{cls} \in \mathbb{R}^{1\times C}$ and the patch tokens $F_p \in \mathbb{R}^{(N-1)\times C}$.

The local branch introduces a convolutional inductive bias to capture fine-grained spatial relationships, operating exclusively on the patch tokens F_p. First, the patch tokens are projected into a lower-dimensional bottleneck representation of dimension C_b. They are reshaped into a 2D feature map X_s. A lightweight multi-scale convolutional block then processes this map. A lightweight multi-scale convolutional block processes this map, comprising parallel depthwise convolutions with different kernel sizes, followed by a pointwise convolution to fuse the resulting features. This operation can be denoted as below:

$$X_s^{'} = \text{Ponv}(\sum_{k\in\{3,5,7\}} \text{Donv}_k(X_s)) + X_s, \tag{5}$$

where Donv_k denotes a depthwise convolution with kernel size k, and Ponv is a 1×1 pointwise convolution.

The resulting feature map is reshaped back into a sequence, passed through a GELU activation, and projected to the original dimension C to generate the patch token residual. The local branch does not modify the class token, so its corresponding residual is zero. The complete local residual can be formulated as below:

$$\Delta F_{local} = [\mathbf{0}^{1\times C}; \text{Linear}(\text{GELU}(\text{Reshape}(X_s^{'})))], \tag{6}$$

where $[\cdot;\cdot]$ denotes concatenation along the sequence dimension. In parallel, the global branch applies a standard adapter architecture to the input tensor F, refining features in a channel-wise manner. This is formulated as below:

$$\Delta F_{global} = \text{Linear}_{up}(\text{GELU}(\text{Linear}_{down}(F))). \tag{7}$$

This structure allows for efficient, low-rank adaptation of the global feature representations for all tokens. The final adaptive residual is the sum of the outputs from both branches. The updated feature tensor F' is computed as:

$$F^{'} = F + \Delta F_{global} + \Delta F_{local}. \tag{8}$$

To ensure training stability and preserve the valuable pre-trained knowledge, the final linear layers in both branches are initialized with zeros. This strategy ensures the modulator initially functions as an identity connection, allowing it to progressively learn a meaningful residual during training.

Decision-Driven Orthogonal Constraint

For any pair of an original image x_p and its corresponding compressed version x_c, after passing through the bidirectional enhanced feature extractor, we obtain the feature vectors $F_p^{'}$ and $F_c^{'}$, respectively. We define the compression perturbation vector v_c as the difference between these two feature vectors:

$$v_c = F_p^{'} - F_c^{'}. \tag{9}$$

Method	InfoMax-GAN	BE-GAN	Cramer-GAN	Att-GAN	MMD-GAN	Rel-GAN	S3-GAN	SNG-GAN	STG-GAN	Pro-GAN	Style-GAN	Style-GAN2	Big-GAN	Cycle-GAN	Star-GAN	Gau-GAN	Deep-fake	Mean Acc
MesoNet (2018)	50.5	50.6	50.0	50.5	50.2	50.4	49.3	50.2	50.9	51.4	51.5	54.2	52.0	53.4	50.6	53.4	50.0	51.2
FF++ (2019)	74.4	30.6	75.5	64.2	76.3	61.5	54.9	71.8	82.2	90.4	60.4	65.2	60.5	80.0	74.4	72.3	51.0	67.3
F3Net (2020)	65.9	42.6	68.9	55.9	63.7	56.3	53.1	62.1	74.9	84.1	56.7	60.4	56.1	77.8	71.2	68.6	50.4	63.2
MAT (2021)	54.5	49.8	59.7	50.1	57.8	50.8	52.8	52.8	56.7	85.7	52.4	53.1	52.9	72.2	57.6	67.6	51.1	57.7
SBIs (2022)	56.6	51.9	63.4	50.1	59.3	50.6	62.2	52.1	53.0	88.4	51.2	52.4	55.4	74.8	53.6	78.3	51.1	59.3
ADD (2022)	52.0	51.0	59.0	50.7	57.2	52.7	44.7	52.3	53.1	70.9	48.0	48.4	51.7	72.4	55.7	64.7	51.3	55.2
QAD (2023)	74.8	53.7	79.6	60.1	78.3	66.5	56.0	76.3	80.4	86.3	55.4	57.2	59.1	77.1	79.9	65.8	55.8	69.2
ODDN (2025)	73.1	42.3	76.1	71.2	75.9	72.5	60.5	75.5	85.0	91.3	64.5	69.4	64.3	80.8	78.0	77.3	54.3	<u>71.4</u>
Ours	86.9	64.9	86.4	54.8	84.9	58.7	74.7	73.7	77.5	94.3	59.3	62.5	75.3	93.9	86.3	92.3	55.9	**75.4**

Table 1: Evaluation results for the 2-class training setting under the quality-aware setting. All comparison results are taken from previous work (Tao et al. 2025). **Bold** and <u>underline</u> indicate the best and second-best performance.

Method	InfoMax-GAN	BE-GAN	Cramer-GAN	Att-GAN	MMD-GAN	Rel-GAN	S3-GAN	SNG-GAN	STG-GAN	Pro-GAN	Style-GAN	Style-GAN2	Big-GAN	Cycle-GAN	Star-GAN	Gau-GAN	Deep-fake	Mean Acc
MesoNet(2018)	46.3	44.3	59.7	60.0	59.8	58.7	47.8	56.3	69.1	55.0	51.0	49.9	53.9	60.5	64.8	49.9	51.1	53.3
FF++(2019)	66.9	37.7	79.4	56.6	77.1	60.5	55.0	69.2	79.8	87.8	55.1	59.8	57.1	79.9	75.6	71.6	52.0	66.0
F3Net(2020)	58.0	48.8	61.9	51.5	59.3	53.2	52.0	54.9	61.1	83.4	52.7	54.9	55.0	73.7	65.9	66.7	52.4	59.3
MAT(2021)	54.2	49.9	59.6	50.5	57.6	51.2	52.1	52.7	57.8	86.1	52.3	53.0	52.7	70.3	58.2	68.0	51.3	57.7
SBIs (2022)	56.6	51.9	63.4	50.1	59.3	50.6	62.2	52.1	53.0	88.6	51.3	52.4	55.7	76.0	53.9	78.1	51.2	59.4
ADD (2022)	51.8	50.9	59.0	50.7	57.1	52.8	45.0	52.3	52.9	70.2	48.0	48.7	51.8	71.9	55.5	65.1	51.3	57.9
QAD (2023)	72.3	55.2	80.0	61.5	78.3	65.5	54.5	76.5	79.2	86.4	56.4	58.0	57.4	82.6	77.8	63.5	56.5	68.3
ODDN (2025)	72.1	44.1	76.8	68.1	76.5	73.3	58.0	75.6	83.5	90.8	61.1	65.9	63.9	83.5	77.0	72.9	55.0	<u>70.7</u>
Ours	86.9	64.1	85.6	53.1	83.5	57.7	73.6	72.7	77.0	93.8	57.1	60.7	73.5	94.2	85.9	92.0	55.2	**74.5**

Table 2: Evaluation results for the 2-class training setting under the quality-agnostic setting.

However, the forgery vector cannot be directly obtained in this manner, as it requires paired real images and their synthetic counterparts, which are generally unavailable. To address this limitation, we define a classification decision axis w_f, which points from the real image centroid to the synthetic image centroid. The projection of images onto this axis determines their categories. During training, the classification decision axis can be formulated as follows:

$$\mathbf{w}_f = \frac{\mathbb{E}_{(x,y)\sim\mathcal{B},y=1}[\Phi(x)] - \mathbb{E}_{(x,y)\sim\mathcal{B},y=0}[\Phi(x)]}{\|\mathbb{E}_{(x,y)\sim\mathcal{B},y=1}[\Phi(x)] - \mathbb{E}_{(x,y)\sim\mathcal{B},y=0}[\Phi(x)]\|_2}. \tag{10}$$

Here, $\mathcal{B}$ denotes a training batch, and $\mathbb{E}[\cdot]$ represents the expectation operator. We normalize w_f to unit length, as we are only concerned with the direction it represents.

To ensure that compression features do not interfere with forgery detection, we enforce orthogonality between the compression perturbation vector v_c and the classification decision axis w_f. This geometric constraint guarantees that the feature perturbation has zero projection along the decision axis. We achieve this by introducing a decision-driven orthogonality loss $\mathcal{L}_{\text{ortho}}$, which minimizes the cosine similarity between v_c and w_f. For a batch containing N paired samples (x_p, x_c), where $v_{c,i}$ denotes the perturbation vector for the i-th pair, the loss is formulated as:

$$\mathcal{L}_{ortho} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\mathbf{v}_{c,i}^{\top}\mathbf{w}_f}{\|\mathbf{v}_{c,i}\|_2\|\mathbf{w}_f\|_2}\right)^2. \tag{11}$$

Finally, we jointly optimize the decision-driven orthogonal loss with a standard classification loss L_{cls}. The total training objective L_{total} is defined as below:

$$L_{total} = L_{cls} + \lambda L_{ortho}, \tag{12}$$

where λ is a hyperparameter set to 0.1. Through joint optimization, the detector learns to build a more robust feature space while improving detection accuracy.

Experiments

Experimental Settings

Training dataset. To ensure a fair comparison with existing state-of-the-art methods, we follow the latest baseline setup of ODDN (Tao et al. 2025) and use ProGAN-generated images from ForenSynths (Wang et al. 2020b) as the training set. To simulate the open-world detection scenario, we randomly sample 20% of the data and apply a fixed JPEG compression operation with a quality factor of 40 to create paired data, consisting of both original and compressed versions. The remaining 80% of the data is left uncompressed and serves as unpaired data. In total, our training set is composed of the 20% paired data and the 80% unpaired data.

Test dataset. Our evaluation is conducted on the ForenSynths (Wang et al. 2020b) and the GANGen-Detection (Tan et al. 2024a) dataset, under both Quality-Aware and Quality-Agnostic settings. The ForenSynths dataset consists of images generated by 8 classic GAN models, while the GANGen-Detection dataset contains images from more recent and diverse generation models. Under the Quality-Aware setting, all test images are compressed using the same JPEG quality factor as applied in the paired training data. The quality-agnostic setting is more challenging and reflects real-world scenarios, where each test image is compressed with a randomly selected JPEG quality factor. This simulates the situation encountered when deploying the detector on online social networks, where the model has to han-

Method	InfoMax-GAN	BE-GAN	Cramer-GAN	Att-GAN	MMD-GAN	Rel-GAN	S3-GAN	SNG-GAN	STG-GAN	Pro-GAN	Style-GAN	Style-GAN2	Big-GAN	Cycle-GAN	Star-GAN	Gau-GAN	Deep-fake	Mean Acc
MesoNet(2018)	49.5	46.2	52.6	51.3	53.0	53.8	50.4	51.8	54.2	53.3	49.6	53.9	55.1	50.9	52.3	51.7	45.0	51.4
FF++(2019)	69.5	26.9	80.3	66.8	79.2	69.9	56.2	75.1	84.4	93.6	62.5	60.8	58.5	80.9	78.5	71.0	52.8	68.7
F3Net(2020)	61.0	41.9	65.8	52.9	63.8	55.5	53.8	59.6	71.5	92.2	76.0	59.1	55.9	57.9	71.8	66.0	52.1	62.4
MAT(2021)	57.9	46.9	64.2	50.8	63.4	52.4	52.1	56.2	61.8	90.8	54.2	53.9	52.4	73.1	61.4	64.8	51.2	59.5
SBIs (2022)	60.2	55.7	74.4	50.2	67.1	54.6	61.4	53.0	57.2	96.0	57.4	53.0	55.4	77.6	60.1	74.9	50.6	62.5
ADD (2022)	51.7	50.7	57.3	51.3	55.9	52.4	45.2	51.2	52.4	73.5	49.9	50.1	52.2	70.7	54.4	66.4	51.2	55.3
QAD (2023)	79.9	37.5	79.5	67.4	76.8	71.7	58.0	79.0	83.5	92.7	64.7	68.7	64.0	81.8	80.3	66.3	52.9	70.9
ODDN (2025)	80.6	38.6	80.7	65.8	78.8	71.1	60.5	76.7	85.8	94.0	67.7	69.9	66.7	84.9	80.5	75.2	54.2	<u>72.6</u>
Ours	84.7	70.1	87.9	54.6	85.8	57.2	74.6	74.6	73.4	99.4	69.6	67.3	74.4	91.5	84.7	93.3	57.9	**76.5**

Table 3: Evaluation results for the 4-class training setting under the quality-aware setting.

Method	InfoMax-GAN	BE-GAN	Cramer-GAN	Att-GAN	MMD-GAN	Rel-GAN	S3-GAN	SNG-GAN	STG-GAN	Pro-GAN	Style-GAN	Style-GAN2	Big-GAN	Cycle-GAN	Star-GAN	Gau-GAN	Deep-fake	Mean Acc
MesoNet(2018)	58.7	45.4	63.5	62.9	62.0	50.2	48.7	58.4	64.1	55.4	52.0	48.1	53.7	63.2	62.0	49.6	51.8	54.3
FF++(2019)	68.9	29.9	82.0	63.3	80.4	67.2	55.5	75.4	82.0	93.0	61.1	59.8	57.9	80.1	78.6	67.3	51.9	67.9
F3Net(2020)	62.0	43.4	65.8	53.2	64.1	56.7	55.4	58.8	67.7	92.5	76.6	62.3	56.8	60.5	71.0	71.3	51.1	63.4
MAT(2021)	52.2	49.3	62.5	50.6	60.3	51.7	53.3	53.9	58.6	92.2	54.4	54.9	54.0	76.5	59.4	68.4	51.0	59.4
SBIs (2022)	61.3	57.4	74.8	50.3	67.5	54.6	61.5	53.2	57.1	95.9	57.2	52.9	55.4	78.3	59.3	74.6	50.7	62.6
ADD (2022)	51.0	50.2	54.4	50.3	53.4	50.7	46.2	50.5	50.9	75.8	51.4	51.6	52.7	72.6	52.3	66.4	50.7	55.0
QAD (2023)	76.7	46.4	79.6	68.5	77.1	73.6	58.3	76.3	81.0	90.2	65.3	71.3	64.6	81.8	77.1	66.7	55.1	71.0
ODDN (2025)	80.4	35.1	81.0	68.7	78.2	74.5	62.2	77.5	81.7	91.7	69.2	70.4	68.0	78.8	73.4	73.8	55.3	<u>72.1</u>
Ours	84.7	69.5	87.0	53.5	84.6	56.1	74.0	73.9	73.4	99.5	65.9	64.3	72.4	91.5	84.0	93.1	56.9	**75.5**

Table 4: Evaluation results for the 4-class training setting under the quality-agnostic setting.

dle various compression levels introduced by different platforms. Accordingly, this setting serves as a critical benchmark for evaluating the model's generalization and robustness in practical deployment.

Evaluation metric. Following standard practice in the field (Tao et al. 2025), we adopt accuracy (Acc) as the primary evaluation metric. Accuracy measures the percentage of real and forged images that are correctly classified among all test samples, thereby reflecting the model's overall performance in a straightforward manner.

Implementation details. All experiments are implemented in PyTorch and conducted on NVIDIA Tesla A100 GPUs for both training and testing. Following prior methods (Tao et al. 2025; Tan et al. 2024a), input images are resized to 256×256. During training, random cropping is applied, while center cropping to 224×224 is used during evaluation. The model is optimized using the SAM optimizer with an initial learning rate of 2e-5, a batch size of 32, and trained for 15 epochs.

Main Results

In line with prior works (Tan et al. 2024a; Tao et al. 2025), we train the proposed framework using ProGAN-generated images under two distinct settings to ensure fair evaluation. The images used include both a 2-class (chair, horse) and a 4-class set (car, cat, chair, horse), which are sourced from ForenSynths (Wang et al. 2020b).

Evaluation under 2-class training setting. Table 1 presents the two-class evaluation results under the **quality-aware** setting, where the compression level of test images matches that used during training. The results show that the proposed method achieves an average accuracy of 75.4%, significantly outperforming all existing baselines. Notably, the gradient reversal-based ODDN method reaches 71.4%, and our method improves upon this by 4.0%. This indicates that avoiding direct removal of compression-related features helps preserve forgery-relevant information. Under known compression conditions, our model not only learns a robust decision boundary but also effectively preserves and leverages discriminative cues through the proposed modules, leading to superior performance.

Due to the difference from the fixed compression ratio used in the training stage, the performance of all models declined under the **quality-agnostic** setting, but this further highlights the superiority of our method. As shown in Table 2, our proposed method achieves an average accuracy of 74.5%, substantially outperforming other methods. This indicates that even when the compression strength is unknown, the geometric constraint imposed by $\mathcal{L}_{ortho}$ still forces the compression perturbation direction to be orthogonal to the decision axis. As a result, our model exhibits intrinsic, structural robustness against arbitrary compression levels, rather than being specialized to specific ones seen during training. Such generalization is difficult for GRL-based or other invariance learning methods to achieve, as they often overfit to the particular compression strengths encountered during training. Under this challenging scenario simulating real-world deployment, our method demonstrates strong generalization, proving its great potential in handling unknown and variable image quality conditions.

Evaluation under 4-class training setting. As shown in Table 3, when the content diversity of the training data increases, our proposed method still achieves a leading av-

Method	InfoMax-GAN	BE-GAN	Cramer-GAN	Att-GAN	MMD-GAN	Rel-GAN	S3-GAN	SNG-GAN	STG-GAN	Pro-GAN	Style-GAN	Style-GAN2	Big-GAN	Cycle-GAN	Star-GAN	Gau-GAN	Deep-fake	Mean Acc
Ours	86.9	64.9	86.4	54.8	84.9	58.7	74.7	73.7	77.5	94.3	59.3	62.5	75.3	93.9	86.3	92.3	55.9	**75.4**
w/o C_d	86.4	69.1	84.7	53.6	83.6	57.6	66.1	75.5	77.0	91.3	56.0	57.0	69.1	85.4	81.8	84.2	51.6	72.3
w/o C_m	83.2	59.6	83.9	51.0	80.9	53.4	68.7	71.5	76.6	92.1	53.0	57.1	70.5	90.2	76.2	87.3	50.2	70.9
w/o C_b	57.9	51.9	61.7	53.1	59.3	56.4	66.3	55.4	59.1	85.9	60.2	62.3	64.1	80.4	68.3	80.4	56.3	63.5

Table 5: Effectiveness of the different components. The decision-driven orthogonal constraint C_d is first removed. From this reduced framework, the adaptive global-local modulator C_m is subsequently removed. Finally, the bidirectional update strategy C_b is further removed.

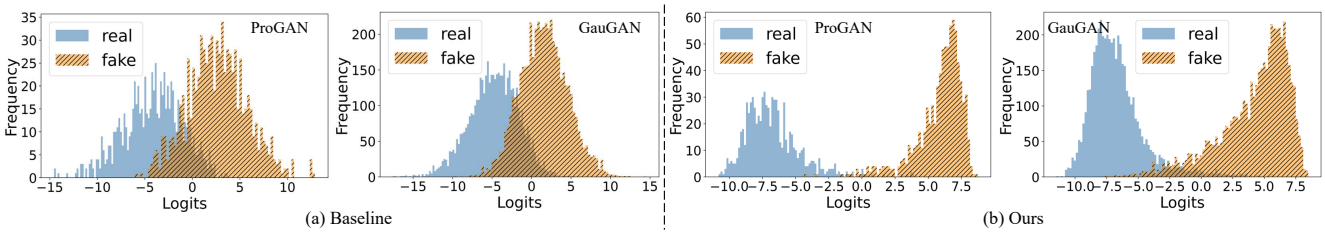


Figure 4: The logit distributions of extracted classification features. Less overlap between features represents a higher degree of separation between them and better generalization.

erage accuracy of 76.5% under the **quality-aware** setting, further demonstrating its strong classification performance and stability. In comparison, ODDN achieves an accuracy of 72.6%. This result indicates that our model does not conflate forgery features with increased content variation.

Under the **quality-agnostic** setting, our method again significantly outperforms all comparison methods with an average accuracy of 75.5%, as illustrated in Table 4. This result confirms that our method is a compression-robust detection framework. It can reliably capture subtle forgery traces even in highly compressed and diverse content scenarios. In summary, across four experimental settings, our proposed method consistently achieves state-of-the-art performance.

Ablation Studies

To validate the effectiveness and necessity of each proposed component, we conduct a comprehensive ablation study, with the results summarized in Table 5. Our full framework achieves a mean accuracy of 75.4%. First, removing the decision-driven orthogonal constraint (C_d) leads to a performance drop of 3.1 points to 72.3%, confirming its essential role in enhancing robustness against compression artifacts. Next, from this reduced baseline, we further remove the adaptive global-local modulator (C_m), resulting in an additional drop of 1.4 points to 70.9%. This indicates that adapting the pre-trained ViT with our modulator provides a distinct and meaningful performance benefit. Finally, removing the bidirectional update strategy (C_b) leads to the most substantial decline, 7.4 points, to 63.5%. This final model essentially represents a plain backbone, and the large performance gap highlights that effective feature fusion between high- and low-frequency streams is the most critical component in our architecture. This sequential analysis clearly illustrates that each component provides a distinct and cumulative contribution, leading to the final state-of-the-art performance of our complete model.

Distribution Visualzation

To provide a qualitative analysis of our model's decision-making process, we visualize the distribution of output logits on the test set (ProGAN and GauGAN) for both a standard baseline and our proposed method, as shown in Figure 4. As seen in Figure (a), the baseline model exhibits significant class confusion. The logit distributions for real (blue) and fake (orange) classes show substantial overlap, indicating that the model produces many low-confidence, indecisive predictions. In contrast, Figure (b) shows that our method achieves clearly improved class separation. The distributions for real and fake samples form two distinct and well-separated clusters, with a clear and wide margin between them. This improved separability provides strong qualitative evidence that our method learns a more discriminative and robust feature representation.

Conclusion

This paper addresses the critical issue of significant performance degradation in AI-generated image detection caused by unknown compression in online social networks. We propose a novel detection framework based on two core principles. First, we introduce a decision-driven orthogonality constraint that geometrically enforces compression artifacts to be orthogonal to forgery artifacts. This approach ensures that the classification decision is not affected by the compression perturbations without removing these compression perturbations, thereby avoiding the loss of discriminative features observed in previous approaches. Second, we design a low- and high-frequency interaction framework to enhance the model's sensitivity to residual high-frequency artifacts while also capturing low-frequency forgery traces. Extensive experiments demonstrate that our method significantly outperforms existing state-of-the-art approaches across multiple benchmarks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grants U24B20175, 62322216, 62311530686), the State Key Laboratory of Public Big Data at Guizhou University (Grant PBD2024-0521), and the Open Research Fund of The State Key Laboratory of Blockchain and Data Security at Zhejiang University.

References

Afchar, D.; Nozick, V.; Yamagishi, J.; and Echizen, I. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. In *WIFS*, 1–7.

Cao, J.; Ma, C.; Yao, T.; Chen, S.; Ding, S.; and Yang, X. 2022. End-to-End Reconstruction-Classification Learning for Face Forgery Detection. In *CVPR*, 4103–4112.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

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 *CVPR*, 7887–7896.

Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; and Holz, T. 2020. Leveraging Frequency Analysis for Deep Fake Image Recognition. In *ICML*, volume 119, 3247–3258.

Ganin, Y.; and Lempitsky, V. S. 2015. Unsupervised Domain Adaptation by Backpropagation. In Bach, F. R.; and Blei, D. M., eds., *ICML*, volume 37, 1180–1189.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; et al. 2014. Generative Adversarial Networks. In *NIPS*, 2672–2680.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *NIPS*, volume 33, 6840–6851.

Kaede, S.; and Toshihiko, Y. 2022. Detecting Deepfakes With Self-Blended Images. In *CVPR*, 18720–18729.

Le, B. M.; and Woo, S. S. 2022. ADD: Frequency Attention and Multi-View based Knowledge Distillation to Detect Low-Quality Compressed Deepfake Images. In *AAAI*.

Le, B. M.; and Woo, S. S. 2023. Quality-Agnostic Deepfake Detection with Intra-model Collaborative Learning. In *ICCV*, 22321–22332.

Li, K.; Ren, W.; Wang, W.; Zhang, L.; and Cao, X. 2025. Detecting Synthetic Image by Cross-Modal Commonality Interaction. In *ACM MM*, 11367–11375.

Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; and Guo, B. 2020. Face X-Ray for More General Face Forgery Detection. In *CVPR*, 5000–5009.

Liu, H.; Tan, Z.; Tan, C.; Wei, Y.; Wang, J.; and Zhao, Y. 2024. Forgery-aware Adaptive Transformer for Generalizable Synthetic Image Detection. In *CVPR*, 10770–10780.

Luo, Y.; Du, J.; Yan, K.; and Ding, S. 2024. LaRE^2: Latent Reconstruction Error Based Method for Diffusion-Generated Image Detection. In *CVPR*.

Ojha, U.; Li, Y.; and Lee, Y. J. 2023. Towards Universal Fake Image Detectors that Generalize Across Generative Models. In *CVPR*, 24480–24489.

Park, N.; and Kim, S. 2022. How Do Vision Transformers Work? In *ICLR*.

Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. In *ECCV*, 86–103.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139, 8748–8763.

Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *ICCV*, 1–11.

Shi, Y.; Gao, Y.; Lai, Y.; Wang, H.; Feng, J.; He, L.; Wan, J.; Chen, C.; Yu, Z.; and Cao, X. 2025. SHIELD: an evaluation benchmark for face spoofing and forgery detection with multimodal large language models. *Vis. Intell.*, 3(1).

Si, C.; Yu, W.; Zhou, P.; Zhou, Y.; Wang, X.; and Yan, S. 2022. Inception Transformer. In *NIPS*.

Sun, W.; Zhou, J.; Lyu, R.; and Zhu, S. 2016. Processing-Aware Privacy-Preserving Photo Sharing over Online Social Networks. In *ACMMM*, 581–585.

Tan, C.; Liu, H.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024a. Rethinking the Up-Sampling Operations in CNN-Based Generative Network for Generalizable Deepfake Detection. In *CVPR*, 28130–28139.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024b. Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Domain Learning. In *AAAI*, 5052–5060.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; and Wei, Y. 2023. Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. In *CVPR*, 12105–12114.

Tao, R.; Le, M.; Tan, C.; Liu, H.; Qin, H.; and Zhao, Y. 2025. ODDN: Addressing Unpaired Data Challenges in Open-World Deepfake Detection on Online Social Networks. In *AAAI*, 799–807.

Wang, H.; Wu, X.; Huang, Z.; and Xing, E. P. 2020a. High-Frequency Component Helps Explain the Generalization of Convolutional Neural Networks. In *CVPR*, 8681–8691.

Wang, S.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020b. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In *CVPR*, 8692–8701.

Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; Hu, H.; Chen, H.; and Li, H. 2023. DIRE for Diffusion-Generated Image Detection. In *ICCV*, 22388–22398.

Wu, H.; Zhou, J.; Tian, J.; Liu, J.; and Qiao, Y. 2022. Robust Image Forgery Detection Against Transmission Over Online Social Networks. *IEEE Trans. Inf. Forensics Secur.*, 17: 443–456.

Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; and Yu, N. 2021. Multi-Attentional Deepfake Detection. In *CVPR*, 2185–2194.