ODDN: Addressing Unpaired Data Challenges in Open-World Deepfake Detection on Online Social Networks

Renshuai Tao[1,2], **Manyi Le**[1,2], **Chuangchuang Tan**[1,2], **Huan Liu**[1,2], **Haotong Qin**[1,3*], **Yao Zhao**[1,2]

[1]Institute of Information Science, Beijing Jiaotong University
[2]Visual Intellgence +X International Cooperation Joint Laboratory of MOE
[3]Center for Project-Based Learning (PBL) D-ITET, ETH Zürich, Switzerland
rstao@bjtu.edu.cn

Abstract

Despite significant advances in deepfake detection, handling varying image quality, especially due to different compressions on online social networks (OSNs), remains challenging. Current methods succeed by leveraging correlations between paired images, whether raw or compressed. However, in open-world scenarios, paired data is scarce, with compressed images readily available but corresponding raw versions difficult to obtain. This imbalance, where unpaired data vastly outnumbers paired data, often leads to reduced detection performance, as existing methods struggle without corresponding raw images. To overcome this issue, we propose a novel approach named the open-world deepfake detection network (ODDN), which comprises two core modules: open-world data aggregation (ODA) and compression-discard gradient correction (CGC). ODA effectively aggregates correlations between compressed and raw samples through both fine-grained and coarse-grained analyses for paired and unpaired data, respectively. CGC incorporates a compression-discard gradient correction to further enhance performance across diverse compression methods in OSN. This technique optimizes the training gradient to ensure the model remains insensitive to compression variations. Extensive experiments conducted on 17 popular deepfake datasets demonstrate the superiority of the ODDN over SOTA baselines.

Code — https://github.com/rstao-bjtu/ODDN/

Introduction

With the rapid development of deep learning-based generation technology(Zhou et al. 2023; Yu et al. 2023; Zhang et al. 2023, 2024a; Pan et al. 2024; Zhang et al. 2024b; Liu, Ye, and Du 2024), AI-generated images are increasingly appearing on social media platforms like Twitter and WeChat. While these images enhance creativity and enjoyment, they also introduce significant safety risks. The ability to create highly realistic images easily has raised concerns about misinformation, privacy, and security. Deepfakes can be used to spread false information, creating fake news that misleads the public. Additionally, these images can be exploited for malicious purposes such as identity theft, fraud, and cyberbullying, amplifying the potential for harm across social media platforms(Asnani et al. 2023; Vice et al. 2024; Ding et al. 2023; Zeng et al. 2024).

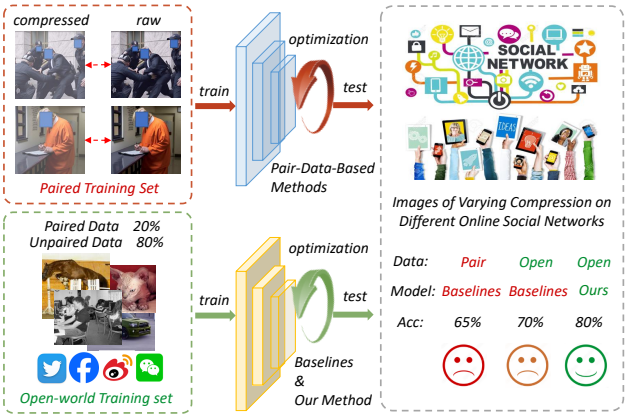

Figure 1: Comparison of training data and models between traditional and open-world scenarios.

Despite significant progress in detecting AI-generated images, detecting forged images on online social networks (OSN) has received relatively little attention. Images on these platforms are often subjected to various compression methods (Dzanic, Shah, and Witherden 2020; Wu et al. 2023), complicating detection efforts. Compression techniques employed by platforms like Twitter and WeChat degrade image quality and obscure manipulation indicators, making forgery identification more challenging. As a result, developing robust detection methods capable of overcoming the unique challenges posed by social media compression remains an urgent research focus in image forensics.

Current methods (Wu et al. 2022; Le and Woo 2023; Liu et al. 2023; Le and Woo 2024) for detecting forged images in compressed formats typically rely on paired data (compressed images and their corresponding originals) for training. These approaches focus on identifying feature correlations between paired data, achieving notable performance improvements under specific compression methods. However, in real-world OSN scenarios, obtaining the original images corresponding to compressed ones is often impractical, leading to a significant imbalance between paired and unpaired data, with unpaired data far outnumbering paired

*Corresponding author.

data. The abundance of unpaired data, whether real or fake, compressed or raw, contains valuable evidence that can aid in authenticity differentiation. However, methods reliant on paired data often fail to effectively integrate this unpaired data, resulting in the loss of critical information. Additionally, existing strategies that focus on fine-grained data associations between compressed images and their corresponding raw versions struggle to address coarse-grained connections among unpaired data. Therefore, exploring robust deepfake detection techniques that can handle open-world scenarios with imbalanced data is crucial for controlling the spread of forged information on various OSN platforms.

In this paper, we first identify the challenge of open-world deepfake detection on OSN, where available training data are often unpaired. Traditional pair-data-based methods are unsuitable for this scenario because compressed images typically lack the corresponding original images, making it difficult to apply these traditional approaches effectively. To address this imbalanced dilemma, we propose a novel method called the open-world deepfake detection network (ODDN), which comprises two core modules: open-world data aggregation (ODA) and compression-discard gradient correction (CGC). The ODA module tackles the challenge of aligning true and false sample features across different types of data, while the CGC module addresses the issue of poor gradient optimization direction when removing compression-sensitive information during training.

Specifically, the ODA module handles paired and unpaired input data from open-world OSN with distinct processing approaches. For a small quantity of paired data (20%), the ODA module exploits fine-grained correlations between the compressed images and their corresponding originals. For the remaining 80% unpaired data, it establishes coarse-grained correlations by clustering the true and false images. Meanwhile, the CGC module ensures the model's insensitivity to compression, which is crucial for effectively handling various compression methods in open-world OSN scenarios. It adopts PCGrad to align and facilitate interactions between distinct gradients, ensuring that the optimization process remains focused on directions that positively impact the main task of the real/fake discrimination.

To comprehensively evaluate the effectiveness of the proposed ODDN, we designed an innovative training data setup to simulate an open-world OSN environment, where unpaired data (80%) far outnumber paired data (20%). Specifically, we compressed a small portion of the training data, typically used for forgery detection tasks, to create paired data, which was then combined with the remaining unpaired data to form the training set. We trained all baseline models and our method on this same training set and assessed their performance under two different test conditions: one aligned with the compression level of the training set and the other unrelated. Our evaluation involved 17 popular GAN-based datasets across these two test settings. The final results demonstrate that our model significantly outperforms existing state-of-the-art models, showcasing its superior effectiveness in OSN. Our contributions are summarized blow: Here's a reconstructed version of the three contributions:

- We introduce the challenge of unpaired data in deepfake detection within open-world scenarios on OSN by designing a novel setup that simulates these environments, offering a valuable benchmark for future research.
- To handle this dilemma, we propose the ODDN, comprising ODA for optimizing artifact feature alignment in unpaired data scenarios, and CGC for reducing gradient biases when removing compression-related information, thereby enhancing detection robustness and adaptability.
- Comprehensive experiments have validated the effectiveness of ODDN across 17 popular datasets under various test settings, demonstrating superior performance in detecting deepfakes on OSN compared to SOTA baselines.

Related Work

Various strategies have been employed to enhance the generalization of detectors to unseen sources. These strategies include diversifying training data through augmentation methods (Wang et al. 2020, 2021), adversarial training (Chen et al. 2022), reconstruction techniques (Cao et al. 2022; He et al. 2021), fingerprint generators (Jeong et al. 2022b), and blending images (Shiohara and Yamasaki 2022). Specific methodologies such as BiHPF (Jeong et al. 2022a) amplify artifacts' magnitudes through two high-pass filters. FreGAN (Jeong et al. 2022c) addresses the overfitting of training sources by mitigating the impact of frequency-level artifacts through frequency-level perturbation maps. Ju et al.(Ju et al. 2022) integrate global spatial information and local informative features in a two-branch model. AltFreezing by Wang et al.(Wang et al. 2023a) leverages both spatial and temporal artifacts for Face Forgery Detection. Approaches by Ojha et al.(Ojha et al. 2023) and Tan et al.(Tan et al. 2023) utilize feature maps and gradients, respectively, as general representations. DIRE by Wang et al. (Wang et al. 2023b) introduces a novel image representation by measuring the feature distance between an input image and its reconstruction counterpart, aiming to alleviate generalization issues.

Method

Problem Definition

Due to the paucity of paired data, specifically an original resolution Deepfake image and its compressed version, the focus of our work is distinct from previous studies. Consequently, to concisely and vividly explain our method, we must adapt the common problem definition used in previous works. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, it comprises two types of images: real images x^r and Deepfake images x^f, with the corresponding labels $y \in \{0, 1\}$ representing real or fake. Subsequently, 20% of the data in $\mathcal{D}$ is randomly chosen to undergo a JPEG compression operation, denoted as P_c, maintaining about 60% image quality. The original versions of these images and the rest of the data in $\mathcal{D}$ are denoted as P and $\tilde{P}$, respectively. Therefore, we can rewrite the composition of the dataset as $\mathcal{D}_{train} = P \cup P_c \cup \tilde{P}$.

As for the inference stage, there are two types: quality-aware and quality-agnostic inference. In the first type, quality-aware inference, the images of the testing set $\mathcal{D}_{test}$

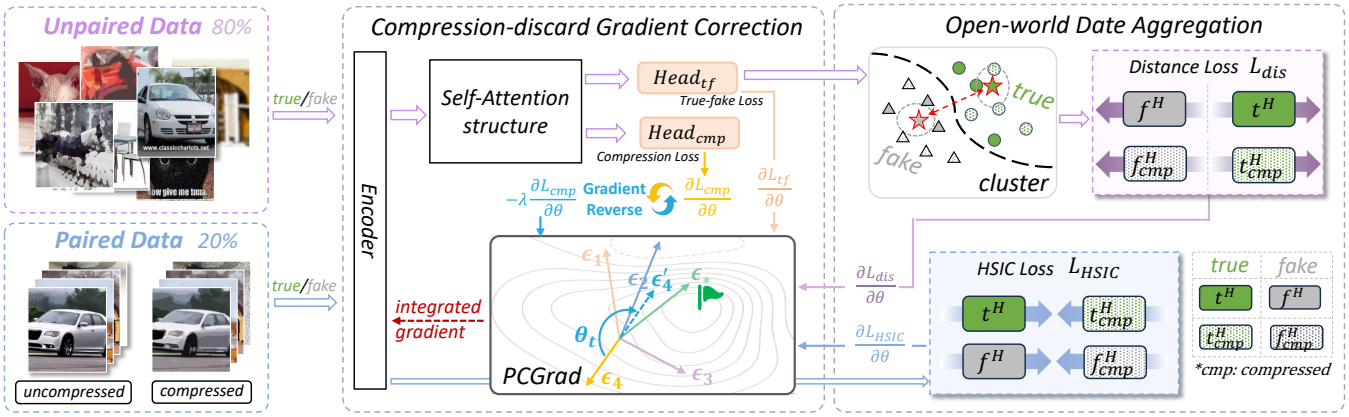

Figure 2: Overview of the proposed Open-world Deepfake Detection Network (ODDN). The ODDN contains two core modules: Open-world Data Aggregation (ODA) and compression-discard Gradient Correction (CGC).

are compressed using the same operation as P_c. In the second type, quality-agnostic inference, the images of the testing set $\mathcal{D}_{test}$ are compressed using various operations to mimic an open-world scenario where the compression type is unknown. The lack of paired data makes it much harder to improve model robustness, creating significant challenges for various efforts, including state-of-the-art approaches.

Network Framework

As illustrated in Figure 2, the proposed ODDN framework comprises two core modules: ODA and CGC. Within the ODA module, data operations are conducted on two types of datasets: 80% unpaired data and other 20% paired data. The unpaired data is aligned using feature-center points, while the paired data is processed using a classic method. The CGC module ensures optimization in the correct direction by integrating gradient correction during the removal of compression-sensitive information. By leveraging a multi-layer adversarial learning mechanism, the framework effectively confounds compression-related characteristics, allowing the detection model to focus on compression-insensitive information. And self-Attention structure is employed to attend to distinct features that are required by different downstream tasks. This approach significantly enhances the generalization capability of deepfake detection models.

Open-world Data Aggregation

The ODA module primarily involves distinct processing of unpaired and paired input data in open-world scenarios.

Solution for the unpaired data. Unlike paired data, unpaired data lacks fine-grained correlations between the compressed and corresponding original images, requiring alternative alignment methods to effectively utilize the abundant unpaired data resources. We aim to establish coarse-grained correlations in unpaired data, which lack strong connections, by clustering true and false images. Specifically, for a given batch of input, we calculate four aggregation centers: the real images C^t, the compressed real images C^t_{cmp}, the fake images C^f, and the compressed fake image C^f_{cmp}. The definition formulas for these four quantities are as follows:

$$C^t = \frac{\Sigma_{i=1}^{N^t} h_i^H}{N^t}, \quad C^t_{cmp} = \frac{\Sigma_{i=1}^{N^t_{cmp}} h_i^H}{N^t_{cmp}} \tag{1}$$

$$C^f = \frac{\Sigma_{i=1}^{N^f} h_i^H}{N^f}, \quad C^f_{cmp} = \frac{\Sigma_{i=1}^{N^f_{cmp}} h_i^H}{N^f_{cmp}} \tag{2}$$

where N^t, N^t_{cmp}, N^f, N^f_{cmp} represent the respective counts of images belonging to four distinct classes within a batch, h_i^H is the feature obtained from the self-attention block.

Subsequently, we enlarge the separation among these cluster centers to improve the distinction between real and fake images, making it easier for the detection model to accurately recognize them. Specifically, to enable our model to effectively classify deepfakes, even in their compressed forms, we strategically increase the distance between the cluster centers of C^t and C^f, as well as the distance S_{cmp} between the compressed clusters C^t_{cmp} and C^f_{cmp}. The detailed formulas are as follows:

$$S = \frac{1}{1 + \Sigma_{i=1}^{d}\sqrt{(C_i^t - C_i^f)^2}} \tag{3}$$

$$S_{cmp} = \frac{1}{1 + \Sigma_{i=1}^{d}\sqrt{(C^t_{i,cmp} - C^f_{i,cmp})^2}} \tag{4}$$

where d denotes the dimension of the hidden features.

The sum of these two types of distances, S_{cmp} and S, can be considered as the alignment loss $\mathcal{L}_{dis}$ for unpaired data. This loss function not only increases the separation between real and fake images but also promotes the aggregation of images within the same class. For greater clarity, the alignment loss of unpaired data can be expressed as follows:

$$\mathcal{L}_{unpair} = S + S_{cmp} \tag{5}$$

Solution for the paired data. Following the previous work (Le and Woo 2023), we observe that the Hilbert-Schmidt Independence Criterion (HSIC), a metric for measuring correlation, is an effective method for maximizing dependency among images of varying quality. This approach

allows the model to learn intricate distribution relationships between paired images. Given the scarcity and value of paired data, we continue to apply HSIC to paired data, as it is a method that can fully exploit the useful information within these pairs. The formula is as follows:

$$\mathcal{L}_{pair} = \widehat{HSIC}(h_c^E,\ h^E) \tag{6}$$

where h_c^E and h^E represent the features of the compressed image and the corresponding original image within the paired data, respectively, as output by the image encoder.

Compression-discard Gradient Correction

This module classifies true and false images using binary cross-entropy loss to distinguish deepfakes, thereby enhancing its ability to detect and identify synthetic content. This process can be formulated as follows:

$$\mathcal{L}_{tf} = \mathcal{L}_{bce}(H_{tf}(h_i^H),\ y_i) \tag{7}$$

where $\mathcal{L}_{bce}$ represents the binary cross-entropy loss, H_{tf} is the head layers of the true/false classification, h_i^H is the feature obtained from the self-attention block, and y_i is the label for the corresponding input sample.

To effectively handle various compression methods in open-world OSN scenarios, the ideal criterion for discrimination should be insensitivity to compression. Thus, we exploit the adversarial learning mechanism, performing effective confusion for discarding the compression information. We assume that compressed images inherently carry a unique signature or fingerprint characteristic of the compression method used, such as JPEG compression. When training on a dataset of compressed images, the model may learn this fingerprint, potentially introducing biases and distorting performance. Our goal is to develop an encoder that maximizes the extraction of features related to fake artifacts while minimizing the inclusion of compression fingerprints. This approach enables the encoder to distinguish between real and fake images, regardless of the compression applied. Inspired by domain-adversarial training of neural networks, we introduce an additional downstream task and use a gradient reversal layer to achieve this goal.

Similarly, the compression-discard loss functions much like the true/false classification branch but differs in data processing. After passing through the image encoder, only the features of paired data are input the compression-discard branch, where they are assessed to determine whether they have been compressed. This operation is defined as follows:

$$\mathcal{L}_{cmp} = \mathcal{L}_{bce}(H_{cmp}(h_i^H),\ y_i) \tag{8}$$

where $y \in \{0, 1\}$, representing compressed or not and H_{cmp} is the head of the compression-discard branch.

Furthermore, the gradient reversal layer inverts the gradient as it passes through. Consequently, when the gradient of $\mathcal{L}_{cmp}$ propagates through the network, the gradients in the encoder and the compression-discard branch have opposite directions but the same magnitude. This operation forces the encoder to discard compression-related information, while the remaining components of the compression classification branch continue to be optimized for detection. The final loss function for the training process is a weighted sum of the above loss functions:

$$\mathcal{L}_{all} = \mathcal{L}_{unpair} + \alpha\mathcal{L}_{pair} + \mathcal{L}_{tf} + \mathcal{L}_{cmp} \tag{9}$$

where α is hyper-parameter that balance the contributions of each component to the overall loss. It is worth noting that GCM effectively leverages valuable information, particularly by utilizing the numerous unpaired data. Additionally, the structure of the branches within GCM is flexible, allowing for the incorporation of other desired models.

However, **with many directions negatively correlated with the gradient direction of the loss $\mathcal{L}_{cmp}$, how can we identify the most suitable direction**? Despite the aforementioned mechanism forcing the encoder to optimize in the reverse gradient direction of $\mathcal{L}_{cmp}$, conflicts often arise between this direction and other gradients. Therefore, it's essential to find a way to align the reverse gradient with other gradients. PCGrad(Yu et al. 2020) offers a solution by projecting conflicting gradients onto the normal vector of another, ensuring constructive interactions among non-conflicting gradients. Inspired by this, we exploit the conflicting gradients projection mechanism to align and facilitate interactions between distinct gradients, ensuring the optimization process remains focused on directions that positively impact the main task. The comprehensive gradient calculation formula is as follows:

$$\nabla_E = \mathbf{P}(\nabla(\mathcal{L}_{pair} + \mathcal{L}_{unpair} + \mathcal{L}_{tf}), -\nabla\mathcal{L}_{cmp}) \tag{10}$$

where ∇_E represents the total gradient calculated for the encoder, ensuring that the gradients are optimized for optimal performance. The symbol **P** denotes the conflicting gradients projection, which is responsible for projecting conflicting gradients onto the normal vector of each other, thereby facilitating interactions among the gradients involved.

Experiments

Settings

Datasets: To ensure a consistent basis for comparison, we employ the training set from *ForenSynths* to train the detectors, in line with the baselines (Wang et al. 2020; Jeong et al. 2022a,c). This training set comprises 20 distinct categories, each featuring 18,000 synthetic images generated using ProGAN, alongside an equal number of real images sourced from the LSUN dataset. For evaluation, we utilized a comprehensive collection of 17 commonly used datasets. The first 8 datasets are derived from the *ForenSynths* (Wang et al. 2020), including images generated by eight distinct generation models. The remaining 9 datasets are derived from the *GANGen-Detection* (Tan and Tao 2024), comprising images generated by nine additional GANs.

Implementation Details: We use the Adam (Kingma and Ba 2015) as the optimizer with a learning rate of 2×10^{-4} and a batch size of 128. For the hyper-parameter α, we adhere to the traditional setting, namely 0.004. In our framework, encoder can be any standard image classifier, such as Res50, to extract features from the image. In order to maintain consistency with previous endeavors(Le and Woo 2023), we employ ResNet-50 (Res50) as our encoder. Our

Method	Info-GAN	BE-GAN	Cram-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
MeNet(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 (**ours**)	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	**71.4**

Table 1: The **quality-aware** experimental results across 17 well-known datasets under the **2-class training data setting**.

Method	Info-GAN	BE-GAN	Cram-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
MeNet(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	71.00	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 (**ours**)	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	**72.6**

Table 2: The **quality-aware** experimental results across 17 well-known datasets under the **4-class training data setting**.

Method	Info-GAN	BE-GAN	Cram-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
MeNet(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 (**ours**)	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	**70.7**

Table 3: The **quality-agnostic** experimental results across 17 well-known datasets under the **2-class training data setting**.

Method	Info-GAN	BE-GAN	Cram-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
MeNet(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 (**ours**)	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	**72.1**

Table 4: The **quality-agnostic** experimental results across 17 well-known datasets under the **4-class training data setting**.

method is implemented using PyTorch on Nvidia GeForce RTX 3090 GPU. We adhere to the commonly used evaluation metrics accuracy (Acc), following common researches.

Quality-aware Experiments

Following the classic setting, we utilize two groups of training sets: a 2-class set ("chair" and "horse") and a 4-class set ("car", "cat", "chair", and "horse") from the *ForenSynths* dataset. The results are presented in Table 1 and 2. To emulate the composition of OSN data in open scenarios, 20% of the data were compressed using operations adopted by popular OSN with constant rate quantization parameters of 40 to create paired data. The remaining 80% were unprocessed to simulate scenarios where unpaired data is significantly more

prevalent than paired data in OSN. It should be clarified that if there are baselines specifically designed for paired data, unpaired data should also be utilized for classification purposes, rather than being left idle, to ensure a fair comparison. During the inference stage, we compress entire images of the testing set by the same compression as the training set and subsequently evaluate each compressed image.

The 2-class and 4-class experimental results presented in Table 1 and 2 compare the performance of various detection methods across 17 different datasets, using the accuracy metric (Acc) as the primary evaluation metric. As shown in Table 1, in the 2-class experiment, the proposed ODDN achieved the highest mean accuracy of 71.4%, significantly outperforming other methods. For instance, QAD, the second-best performer, achieved a mean accuracy of 69.2%, while FF++ and F3Net had mean accuracies of 67.3% and 63.2%, respectively. The proposed method showed particularly strong performance with specific GANs such as ProGAN, STGGAN, and CycleGAN, achieving accuracies of 91.3%, 85.0%, and 80.8%, respectively. This indicates that the method is highly effective in distinguishing between real and fake images in a binary classification setup. Moreover, the method consistently performed well across most datasets, achieving over 70% accuracy in 10 out of the 17. This consistency across various datasets shows the robustness and reliability of the ODDN in quality-aware scenarios.

As shown in Table 2, In the 4-class quality-aware experiment, the proposed method again demonstrated superior performance with the highest mean accuracy of 72.6%, further confirming its effectiveness in more complex classification tasks. QAD followed closely with a mean accuracy of 70.9%, maintaining its position as a strong competitor. Other methods like FF++ and F3Net achieved mean accuracies of 68.7% and 62.4%, respectively, indicating a noticeable performance gap between these methods and the top performers. The proposed method excelled in detecting images from GANs like ProGAN, STGGAN, and CycleGAN, with accuracies of 94.0%, 85.8%, and 84.9%, respectively. These high accuracies highlight the deepfake detection capability of the proposed DANN in another training data scenario. Additionally, DANN showed consistent high performance across most datasets, achieving over 70% accuracy in 11 out of the 17 well-known datasets. This consistency and robustness make it a reliable choice for quality-aware analysis of the generated images in binary classification tasks.

Quality-agnostic Experiments

In this group of quality-agnostic experiments, the training settings are the same with the above quality-aware experiments, that is 2-class and 4-class. It should be noted that the test images used here do not follow the compression applied to the training data. Instead, they are compressed using JPEG compression coefficients sampled from a normal distribution ranging from 30 to 100, simulating open-world scenarios that need to handling unknown compression methods. The evaluation results of the 2-class and 4-class quality-agnostic experiments are shown in Table 3 and 4.

In the 2-class quality-agnostic experiment, the proposed ODDN demonstrated remarkable performance with the highest mean accuracy of 70.7%. This indicates its robustness in identifying real versus fake images without accounting for the quality of the generated images. Specifically, it excelled with GANs such as ProGAN (90.8% accuracy), STGGAN (83.5%), and CycleGAN (83.5%). This high level of performance across diverse GANs underscores the method's adaptability and effectiveness. QAD was the second-best performer with a mean accuracy of 68.3%, making it a reliable alternative but still falling short of the proposed method's overall effectiveness. Other methods like FF++ and F3Net had mean accuracies of 66.0% and 59.3%, respectively, indicating a significant performance gap between these methods and the top performers. These results suggest that while multiple methods are viable for quality-agnostic GAN detection, the proposed ODDN stands out for its consistent and superior performance across baselines.

In the 4-class quality-agnostic experiment, the proposed ODDN again outperformed others with the highest mean accuracy of 72.1%. It achieved notable accuracies with GANs such as ProGAN (91.7%), STGGAN (81.7%), and CycleGAN (78.8%), further demonstrating its robustness in handling more complex classification tasks. QAD followed closely with a mean accuracy of 71.0%, reinforcing its reliability but still trailing behind the proposed method. FF++ and F3Net had mean accuracies of 67.9% and 63.4%, respectively, which, while respectable, highlight the superior consistency and accuracy of the proposed method.

As illustrated in Fig. 3, these results underscore the ODDN's ability to deliver high performance consistently, making it a highly effective choice for quality-agnostic analysis of deepfakes. The robustness across 17 different well-known datasets in different training settings suggests its potential for practical applications in open-world scenarios where distinguishing GAN-generated content is crucial.

Ablation Study

The ablation study presented in the table evaluates the impact of different components (ODA and CGC) on the performance of the baseline method across the 17 datasets, using mean accuracy (Acc) as the metric. The results are analyzed in three configurations: the baseline, with ODA, and both.

The baseline achieved a mean accuracy of 69.4%, showing strong performance across several GANs. Notably, it performed exceptionally well with ProGAN (90.8%), STGGAN (84.8%), and CycleGAN (81.8%). However, there were GANs where the baseline method's performance was less impressive, such as AttGAN (66.9%) and BEGAN (40.9%). Adding ODA to the baseline resulted in an improvement in mean accuracy, increasing to 71.0%. This enhancement indicates that ODA positively contributes to the model's ability to distinguish between real and fake images. Specific datasets like InfoMaxGAN, MMDGAN, and StyleGAN2 saw noticeable improvements, with accuracy increasing to 75.2%, 78.9%, and 62.8%, respectively. The improvements were consistent across most datasets, demonstrating the robustness of the ODA component. Further adding CGC to the baseline with ODA configuration led to the highest mean accuracy of 71.4%. This configuration achieved the best performance across datasets, indicating that the

Method	Info-GAN	BE-GAN	Cram-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
Baseline	74.1	40.9	78.9	66.9	77.0	70.6	57.4	77.3	84.8	90.8	58.8	61.4	60.1	81.8	79.7	70.6	53.1	69.4
+ ODA	75.2	39.0	80.2	63.9	78.9	73.3	59.7	77.2	84.3	92.8	62.8	62.7	62.3	83.3	80.0	78.1	54.1	71.0
+ CGC (**ours**)	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	**71.4**

Table 5: The experimental results of the ablation study. The settings are the same as 2-class quality-aware experiment above.

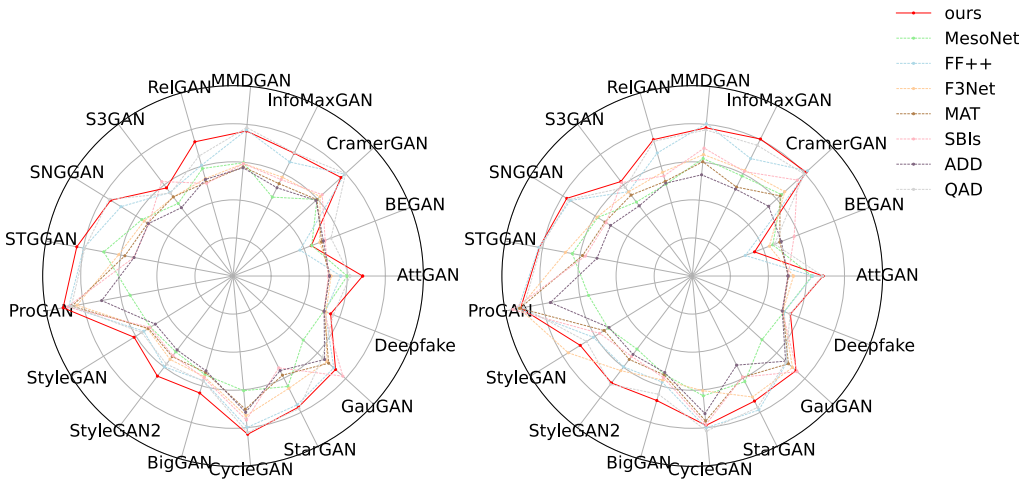

Figure 3: Performance comparison across 17 well-known datasets in quality-agnostic experiments (simulating the open-world OSN scenario) is illustrated for both the 2-class (left figure) and 4-class (right figure) training settings.

combination of ODA and CGC significantly enhances the model's overall accuracy. The method excelled particularly with ProGAN (91.3%), STGGAN (85.0%), and CycleGAN (80.8%). The combined approach also improved performance in datasets where the baseline method had lower accuracy, such as BE-GAN and Cramer-GAN, demonstrating its effectiveness in a broader range of scenarios.

Feature Distribution Visualization

To verify the consistency of invariant representation across input quality, we visualized the feature distribution of the baseline and ODDN using t-SNE (Van der Maaten and Hinton 2008) in 3D, observing from six angles: front, rear, right, left, up, and down (Figure 4). For the baseline model, compressed deepfake features tend to cluster closely with other features, making them difficult to distinguish from multiple perspectives. This close proximity is likely a key reason for the baseline's reduced detection performance. In contrast, ODDN significantly increases the separation between features of different classes, with each class clearly occupying distinct regions in at least one of the six observed directions. This greater separation allows for more effective distinction between features, leading to improved detection accuracy. In summary, the proposed ODDN demonstrates superior generalization across varying input qualities, enhancing its performance in distinguishing deepfakes.

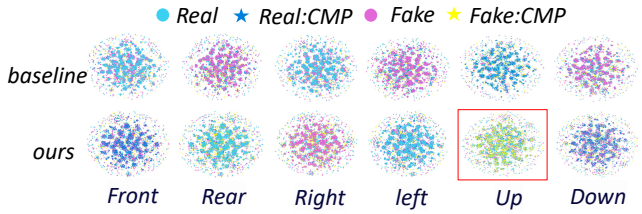

Figure 4: The feature visualization of baseline and ODDN.

Conclusion

In conclusion, this paper presents the ODDN, a novel approach designed to address the challenges of deepfake detection in open-world scenarios, particularly on online social networks where unpaired data is prevalent. Through the integration of two key modules: ODA and CGC, ODDN effectively handles the complexities associated with varying data qualities and compression methods. The comprehensive experiments are conducted across 17 popular datasets under diverse test settings demonstrate that ODDN significantly outperforms existing SOTA models.

Acknowledgments

This work was supported by Beijing NSF (No.L242021, No.4222014), National NSF of China (No.U24B20179, No.62336001), and the Talent Fund of Beijing Jiaotong University (No.2024XKRC047, No.2024XKRC011).

References

Afchar, D.; Nozick, V.; Yamagishi, J.; and Echizen, I. 2018. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, 1–7. IEEE.

Asnani, V.; Yin, X.; Hassner, T.; and Liu, X. 2023. Reverse engineering of generative models: Inferring model hyperparameters from generated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cao, J.; et al. 2022. End-to-end reconstruction-classification learning for face forgery detection. In *CVPR*, 4113–4122.

Chen, L.; et al. 2022. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *CVPR*, 18710–18719.

Ding, H.; Sun, Y.; Huang, N.; Shen, Z.; and Cui, X. 2023. TMG-GAN: Generative Adversarial Networks-Based Imbalanced Learning for Network Intrusion Detection. *IEEE Transactions on Information Forensics and Security*, 19: 1156–1167.

Dzanic, T.; Shah, K.; and Witherden, F. 2020. Fourier spectrum discrepancies in deep network generated images. *Advances in neural information processing systems*, 33: 3022–3032.

He, Y.; et al. 2021. Beyond the Spectrum: Detecting Deepfakes via Re-Synthesis. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 2534–2541. International Joint Conferences on Artificial Intelligence Organization.

Jeong, Y.; et al. 2022a. BiHPF: Bilateral High-Pass Filters for Robust Deepfake Detection. In *WACV*, 48–57.

Jeong, Y.; et al. 2022b. FingerprintNet: Synthesized Fingerprints for Generated Image Detection. In *ECCV*, 76–94. Springer.

Jeong, Y.; et al. 2022c. FrePGAN: robust deepfake detection using frequency-level perturbations. In *AAAI*, volume 36, 1060–1068.

Ju, Y.; et al. 2022. Fusing Global and Local Features for Generalized AI-Synthesized Image Detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, 3465–3469. IEEE.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.

Le, B. M.; and Woo, S. S. 2023. Quality-agnostic deepfake detection with intra-model collaborative learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22378–22389.

Le, B. M.; and Woo, S. S. 2024. Gradient alignment for cross-domain face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 188–199.

Liu, F.; Ye, M.; and Du, B. 2024. Learning a generalizable re-identification model from unlabelled data with domain-agnostic expert. *Visual Intelligence*, 2(1): 28.

Liu, J.; Zhou, J.; Wu, H.; Sun, W.; and Tian, J. 2023. Generating Robust Adversarial Examples against Online Social Networks (OSNs). *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(4): 1–26.

Ojha, U.; et al. 2023. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 24480–24489.

Pan, S.; Zhang, Z.; Wei, K.; Yang, X.; and Deng, C. 2024. Few-shot Generative Model Adaptation via Style-Guided Prompt. *IEEE Transactions on Multimedia*.

Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, 86–103. Springer.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1–11.

Shiohara, K.; and Yamasaki, T. 2022. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18720–18729.

Tan, C.; and Tao, R. 2024. GANGen-Detection: A Dataset generated by GANs for Generalizable deepfake Detection. https://github.com/chuangchuangtan/GANGen-Detection.

Tan, C.; et al. 2023. Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. In *CVPR (CVPR)*, 12105–12114.

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

Vice, J.; Akhtar, N.; Hartley, R.; and Mian, A. 2024. Bagm: A backdoor attack for manipulating text-to-image generative models. *IEEE Transactions on Information Forensics and Security*.

Wang, C.; et al. 2021. Representative forgery mining for fake face detection. In *CVPR*, 14923–14932.

Wang, S.-Y.; et al. 2020. CNN-generated images are surprisingly easy to spot... for now. In *CVPR*, 8695–8704.

Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; and Li, H. 2023a. AltFreezing for More General Video Face Forgery Detection. In *CVPR*, 4129–4138.

Wang, Z.; et al. 2023b. DIRE for Diffusion-Generated Image Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 22445–22455.

Woo, S.; et al. 2022. Add: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 122–130.

Wu, H.; Zhou, J.; Tian, J.; Liu, J.; and Qiao, Y. 2022. Robust image forgery detection against transmission over online social networks. *IEEE Transactions on Information Forensics and Security*, 17: 443–456.

Wu, H.; Zhou, J.; Zhang, X.; Tian, J.; and Sun, W. 2023. Robust Camera Model Identification over Online Social Network Shared Images via Multi-Scenario Learning. *IEEE Transactions on Information Forensics and Security*.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33: 5824–5836.

Yu, Y.; Yang, W.; Ding, W.; and Zhou, J. 2023. Reinforcement learning solution for cyber-physical systems security against replay attacks. *IEEE Transactions on Information Forensics and Security*.

Zeng, K.; Chen, K.; Zhang, J.; Zhang, W.; and Yu, N. 2024. Towards Secure and Robust Steganography for Black-box Generated Images. *IEEE Transactions on Information Forensics and Security*.

Zhang, C.; Ming, Y.; Wang, M.; Guo, Y.; and Jia, X. 2023. Encrypted and compressed key-value store with pattern-analysis security in cloud systems. *IEEE Transactions on Information Forensics and Security*.

Zhang, S.; Yang, Y.; Zhou, Z.; Sun, Z.; and Lin, Y. 2024a. DIBAD: A Disentangled Information Bottleneck Adversarial Defense Method using Hilbert-Schmidt Independence Criterion for Spectrum Security. *IEEE Transactions on Information Forensics and Security*.

Zhang, Z.; Pan, S.; Wei, K.; Ji, J.; Yang, X.; and Deng, C. 2024b. Few-Shot Generative Model Adaption via Optimal Kernel Modulation. *IEEE Transactions on Circuits and Systems for Video Technology*.

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*, 2185–2194.

Zhou, Z.; Dong, X.; Meng, R.; Wang, M.; Yan, H.; Yu, K.; and Choo, K.-K. R. 2023. Generative steganography via auto-generation of semantic object contours. *IEEE Transactions on Information Forensics and Security*.