# Unlocking the Potential of Lightweight Quantized Models for Deepfake Detection

**Renshuai Tao**[1,2] , **Ziheng Qin**[1,2] , **Yifu Ding**[3*] , **Chuangchuang Tan**[1,2] ,
**Jiakai Wang**[1] and **Wei Wang**[1,2]
[1]Institute of Information Science, Beijing Jiaotong University
[2]Visual Intellgence +X International Cooperation Joint Laboratory of MOE
[3]School of Computer Science and Engineering, Beihang Univeristy
{rstao, zihengqin, tanchuangchuang, wei.wang}@bjtu.edu.cn, yifuding@buaa.edu.cn,
wangjk@mail.zgclab.edu.cn

## Abstract

Deepfake detection is increasingly crucial due to the rapid rise of AI-generated content. Existing methods achieve high performance relying on computationally intensive large models, making real-time detection on resource-constrained edge devices challenging. Given that deepfake detection is a binary classification task, there is potential for model compression and acceleration. In this paper, we propose a low-bit quantization framework for lightweight and efficient deepfake detection. The Connected Quantized Block extracts common forgery features via the quantized path and retains method-specific textures through the shortcut connections. Additionally, the Shifted Logarithmic Redistribution Quantizer mitigates information loss in near-zero domains by unfolding the unbalanced activations, enabling finer quantization granularity. Comprehensive experiments demonstrate that this new framework significantly reduces **10.8**× computational costs and **12.4**× storage requirements while maintaining high detection performance, even surpassing SOTA methods using less than **5%** FLOPs, paving the way for efficient deepfake detection in resource-limited scenarios. Code is in https://github.com/rstao-bjtu/QMDD.

## 1 Introduction

With the advancement of deep learning [Tao *et al.*, 2021; Wei *et al.*, 2020; Wu *et al.*, 2024; Tao *et al.*, 2022], recent progress in deepfake detection [Liu *et al.*, 2023; Tan *et al.*, 2025; Le and Woo, 2024; Tao *et al.*, 2025a] has increasingly relied on large-scale models, including convolutional neural networks (CNNs) and transformer-based architectures, to achieve state-of-the-art performance. These models [Tao *et al.*, 2025b; Le and Woo, 2024] excel at capturing intricate patterns within data, allowing them to detect even the most subtle manipulations. However, their reliance on computational resources for both training and inference often brings difficulties in deployment on resource-constrained environments, such as edge applications or real-time scenarios [Asnani *et al.*, 2023; Vice *et al.*, 2024; Ding *et al.*, 2023; Zeng *et al.*, 2024]. The rise of AI-generated content on social platforms [Wu *et al.*, 2023] necessitates the development of efficient detection systems to have low latency and computational requirements, no matter deploying on the cloud to process vast volumes of data, or carrying on the edge devices to safeguard the end users. Addressing the challenge requires minimizing the size and computational complexity of deepfake detection models without sacrificing performance. A promising approach is model quantization technique, which reduces the data precision and compacts the model while keeps the accuracy , paving the way for the deployment of deep learning models on resource-constraint scenarios.

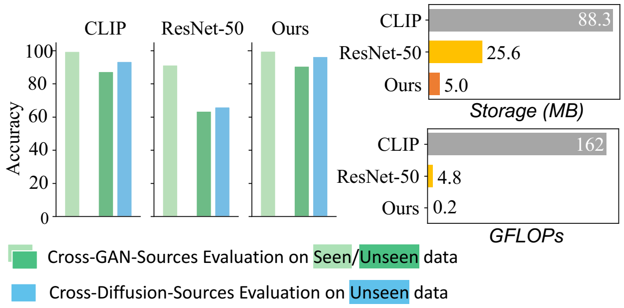


Figure 1: Comparison of deepfake detection models with different models, including the accuracy performance in four cases, the model storage and computational complexity.

The simplicity of the binary classification task suggests possible redundancy , leaving room for compression. As illustrated in Figure 1, we compare CLIP-ViT-L with 124M parameters, ResNet-50 with only 25.6M parameters and our model across three testing scenarios: seen domains generated by the same generator as the training data (light green) and unseen domains generated by other GAN generators (dark green) and Diffusion models (blue). The figure shows that smaller models can also do the same job. This raises the question: ***Do we really need to rely on large models for deepfake detection, especially on resource-constrained edge devices?***

In this paper, we propose a novel quantization framework to compress deepfake detection models while maintaining high performance. We first identify and validate the effectiveness of the NPR (Neighboring Pixel Relationships) rep-

[*] Corresponding author.

resentation as a quantization-friendly alternative compared to original RGB images, which displays better compatibility to quantization errors, ensuring the feature consistency with low-bit representations. Based on this, we design a quantized model equipped with the **Connected Quantized Block (CQB)**, which extracts forgery features by combining method-specific textures from full-precision shortcut connections with common features captured in the quantized main path. The auxiliary connections in CQB help preserve fine-grained details, effectively addressing the distortion caused by quantization. Additionally, we introduce the **Shifted Logarithmic Redistribution (SLR)** quantizer to handle the unbalanced distribution of intermediate activations. By unfolding the negative ranges, SLR enhances the representation of forgery-related features and reduces information loss resulting from coarse quantization granularity. Together, these innovations form a robust framework for efficient and accurate deepfake detection.

Our carefully designed quantization framework achieves a substantial reduction in computational overhead by 10.8× and model storage by 12.4× without compromising performance. Notably, it outperforms state-of-the-art deepfake detection methods such as CNNDetection [Wang and others, 2020], LGrad [Tan *et al.*, 2023], FrePGAN [Jeong and others, 2022c] using less than $5\%$ of the FLOPs. The results validate the hypothesis of parameter redundancy in deepfake detection models and strongly demonstrates the potential of lightweight models. Beyond reducing memory footprint and inference latency, our framework enhances the ability to extract forgery-specific features across various generators, paving the way for efficient and scalable deepfake detection systems, especially for resource-constrained environments. The main contributions are summarized as follows:

- We are the **first to explore quantization in deepfake detection**. Through extensive experiments, we demonstrate the potential of smaller models.
- We propose the Connected Quantized Block to complement method-specific fine textures into universal artifacts , compensating for the details.
- We introduce the Shifted Logarithmic Redistribution Quantizer, unfolding the clustered and unbalanced activations to mitigate the loss from coarse quantization.
- Our quantized deepfake detection model achieves a remarkable 10.8× and 12.4× reduction in computation and model storage, while maintaining or even surpassing the accuracy of full-precision counterparts.

## 2 Related Work

### 2.1 Deepfake Detection

Various strategies have been employed to enhance the generalization of detectors to unseen sources. These strategies include diversifying training data through augmentation methods [Wang and others, 2020; Wang and others, 2021; Yin *et al.*, 2024], adversarial training [Chen and others, 2022], reconstruction techniques [Cao and others, 2022; He and others, 2021], fingerprint generators [Jeong and others, 2022b], and blending images [Shiohara and others, 2022]. Specific methodologies such as BiHPF [Jeong and others, 2022a] amplify artifacts' magnitudes through two high-pass filters. FreGAN [Jeong and others, 2022c] addresses the overfitting of training sources by mitigating the impact of frequency-level artifacts through frequency-level perturbation maps. Ju et al.[Ju and others, 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 Deepfake Detection. Approaches by Ojha et al.[Ojha *et al.*, 2023] and Tan et al.[Tan and others, 2023] utilize feature maps and gradients, respectively, as general representations. DIRE by Wang et al. [Wang and others, 2023] introduces a novel image representation by measuring the feature distance between an input image and reconstruction counterpart, aiming to alleviate generalization issues.

### 2.2 Model Quantization

Quantization has been applied on various downstream applications as a widely-used compression technique, due to its practicability. Particularly in platforms with limited computational resources and low task complexity, such as monitoring camera, mobile apps, and IoT devices [Ding *et al.*, 2024; Hernández *et al.*, 2024; Wang *et al.*, 2023b]. However, few works focus on deepfake detection. Lanzino et al. [Lanzino *et al.*, 2024] utilizes binarization on deepfake detection to uncover manipulation traces in frequency and texture domains but reduce the computational overhead by binarized operations. However, since binarization is an extreme quantization technique, the accuracy drop is always inevitable, and the computational complexity and parameter storage can be further reduced, which we have compared in our experiments.

## 3 Method

### 3.1 Problem Definition

A widely adopted approach for deepfake detection involves training a binary classifier using supervised deep learning methods to differentiate between real and fake images. This task is predominantly deployed on the cloud due to the limited computational capabilities of user-end devices. With the rise in popularity of video streaming platforms, deploying deepfake detection models directly on edge devices can greatly lower detection costs and latency while improving user security by protecting against deception from AI-generated content. To enable deployment on edge devices, minimizing the model's computational cost and storage size is crucial. However, the impact of backbone network size on this task remains unexplored. To address this, we evaluate CNNDetection [Wang and others, 2020] with backbones of varying sizes, assessing its performance on the training data domain as well as its generalization capabilities to other domains generated by different models. As shown in Table 1, increasing the backbone network size does not yield significant improvements in generalization performance. This phenomenon has led us to consider: Do we truly need larger models to perform binary classification for deepfake detection tasks, particularly in resource-limited scenarios?

| | ResNet-23 | ResNet-34 | ResNet-50 |
|---|---|---|---|
| Seen | 99.9% | 99.9% | 99.9% |
| Unseen | 77.6% | 80.0% | 79.4% |

Table 1: Performance of different backbones on ForenSynths.

***Opportunities and Challenges of Quantization.*** We apply quantization techniques to deepfake detection models and achieve promising results. Specifically, we quantize a ResNet-23 model to 8-bit, 4-bit, and 2-bit using LSQ [Esser *et al.*, 2019], and further binarize it using ReActNet [Liu *et al.*, 2020]. Results in Figure 2 reveal that the quantized models maintain comparable accuracy, especially when tested on the seen data domain generated by ProGAN. It suggests that deepfake detection models have redundancy in parameter bitwidth, leaving room for further compression by quantization techniques.

## 3.2 Preliminary: Quantization-friendly NPR Representation

To harness the benefits of quantization while addressing the issue of detail loss caused by rounding, we propose a tailored quantization framework for deepfake detection.

Specifically, we adopt NPR [Tan *et al.*, 2024b] (Neighboring Pixel Relationships) as the representation of artifacts due to its robustness against quantization. NPR is an effective approach for detecting forged images by capturing upsampling artifacts, i.e., patterns generated by interpolation or transposed convolutions, which are commonly used in generative models. The definition of NPR is

$$\hat{x} = \mathrm{up}(x), \tag{1}$$

$$I = \mathrm{conv}(\hat{x}), \tag{2}$$

where $x$ is the up-scaled feature map. Therefore, NPR focuses on the relationships between neighboring pixels, highlighting that these local interactions are more likely to contain potential artifacts. By removing semantic information while preserving edges, patterns, and pixel relationships, NPR reduces the reliance on large receptive fields, enabling the network to distinguish between real and forged images by focusing on small-scale details and artifact patterns.

Generally, NPR smooths out minor differences and thus enhances contrasts, which is an fundamentally similar effect to quantization. It suggests us to verify the compatibility between NPR and quantization, particularly in preserving forgery-related details.

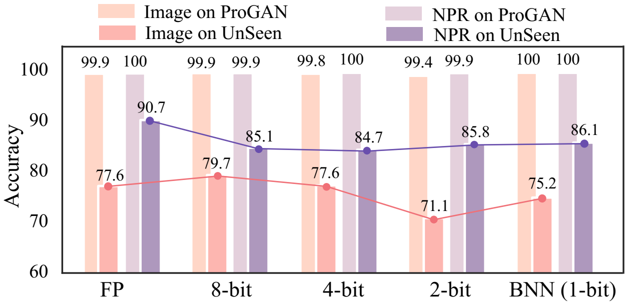


Figure 2: Effect of quantization to various bitwidths on ResNet-23.

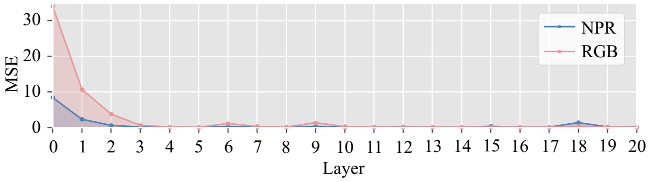
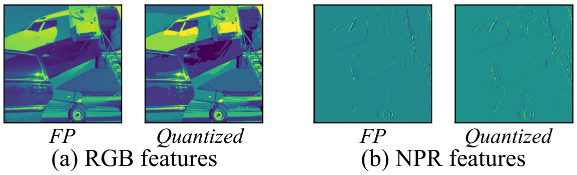


Figure 3: Quantization error of each layer when the input data is original RGB images or NPR representations.



(a) RGB features (b) NPR features

Figure 4: Visualizations for (a) RGB images and (b) NPR-processed representations of full-precision (*FP*) and MinMax *Quantized*.

First, quantization maps continuous real numbers to a discrete space with $N$ values ($N = 2^k$, $k$ is the bitwidth). For example, in integer quantization, we have predefined threshold set $T$, and the input data $x$ can be quantized as

$$Q(x) = \sum_{i=0}^{N-1} \mathbb{I}_i [\![ x \geq t_i ]\!], \tag{3}$$

where $t_i = \frac{1+2(i-1)}{2^k-1} \in T^N$ is the threshold of $i$-th value and $\mathbb{I} = \{1, \dots, 1\} \in Z^N$ is an all-one vector. $[\![\cdot]\!]$ is the Kronecker delta function, which takes the value of 1 if the condition is met; otherwise, it takes 0. Methods like LSQ [Esser *et al.*, 2020], N2UQ [Liu *et al.*, 2022] and QIL [Jung *et al.*, 2019] have a trainable threshold set $T$, allowing the model to learn optimal quantization parameters during training.

Since the representation space is limited, it is inevitable for quantization to smooth out fine details, which can be formulated as

$$Q(x \pm \delta) - Q(x) = 0, \tag{4}$$

where $\delta$ represents a small variation in the image feature adjacent to the pixel $x$, as the quantization granularity around $x$ is greater than $\delta$. Fortunately, when applying $\mathrm{NPR}(\cdot)$ representation as pre-processing, the quantization within the adjacent variation of $x$ can be

$$Q(\mathrm{NPR}(x \pm \delta)) - Q(\mathrm{NPR}(x)) = Q(\delta) - Q(0). \tag{5}$$

It shows the consistency of NPR and quantization in terms of smoothing minor changes in adjacent pixels. Furthermore, NPR concentrates the image into a limited color space, only preserving the outline of the figure. It facilitates quantization by finer-grained threshold set $T$ to minimize quantization error and retain essential details.

We perform a quantitative analysis of the quantization error by calculating the MSE loss for each layer using the images and NPR representations of Motorbike class in ProGAN dataset. The results are plotted in Figure 3. It suggests that

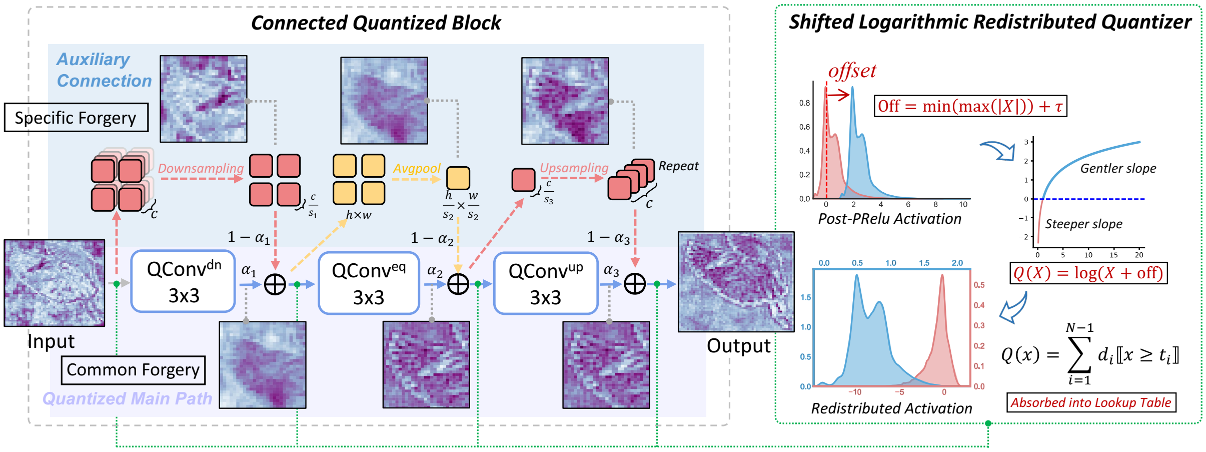


Figure 5: Overview of our quantization framework for the deepfake detection task.

using the original image as input leads to larger quantization errors, particularly in the earlier layers. Although the quantization errors diminish in the later layers, the fine-grained features may already be disrupted.

To give an intuitive conception, we visualize the RGB image and NPR representation under full-precision and 4-bit quantization. As shown in Figure 4, RGB images under 4-bit quantization experience noticeable changes in lighting and pixel intensity relationships, resulting in sharper visual contrasts. This quantization process alters the balance of brightness and shadows. These effects are likely to be misinterpreted as artifacts generated by AI models during deepfake detection. However, quantization applied to NPR preserves most of the information.

### 3.3 Framework Overview

As shown in the left part of Figure 5, our framework introduces a lightweight CNN module named the **Connected Quantized Block (CQB)**, built upon residual structures. It consists of a quantized main path with basic CNN blocks and a full-precision auxiliary path using only pooling and sampling. The quantized path, trained with learnable low-bit quantization, captures general features, albeit with potential detail loss due to rounding. To compensate, the auxiliary path preserves fine textures via shortcut connections with minimal overhead, effectively mitigating quantization-induced information loss. Details are provided in Sec. 3.4. To further address activation imbalance, where values cluster densely in certain regions, we propose **Shifted Logarithmic Redistribution (SLR)** (see the right of Figure 5). SLR redistributes compact negative activations, reducing quantization errors caused by uniform binning and recovering critical texture details for forgery detection. Its formulation is discussed in Sec. 3.5.

### 3.4 Connected Quantized Block for Forgery Feature Compensation

In this section, we first outline the construction of the quantized CNN block and demonstrates the detailed implementation of the shortcut connections.

When quantized to extremely low bit-widths, the representation capacity of the parameters becomes significantly constrained. For instance, in a convolution layer $\mathrm{Conv}(\cdot)$ with $k$-bit, the integer space is limited to $2^k$ discrete values. As $k$ decreases, this discrete space becomes smaller, leading to restricted representation capabilities. To mitigate the limit, we enhance the original block with a few full-precision paths, preserving and restoring the lost details with minimal computational cost. This approach is hardware-friendly and efficient, as it does not introduce additional intensive computations (such as multiplications) in the auxiliary path.

Specifically, we introduce the **Connected Quantized Block (CQB)** based on the residual block. The expand and reduction in the full-precision path introduced by CQB can be flexibly adjusted through the scaling ratio $s$. A typical residual block comprises three types of quantized convolution layers: $\mathrm{QConv}^{\mathrm{dn}}$ for downsampling, $\mathrm{Conv}^{\mathrm{up}}$ for upsampling and $\mathrm{Conv}^{\mathrm{eq}}$ for those do not change the channel number, respectively. In our CQB, these layers are all quantized, and we add weighted full-precision shortcut connections besides the three layers. For downsampling layer $\mathrm{Conv}^{\mathrm{dn}}$, we take the average for every $s$ input:

$$\mathrm{QConv}^{\mathrm{dn}}(x_i) = \alpha_1 Q(\mathrm{Conv}^{\mathrm{dn}}(x_i)) \tag{6}$$

$$+(1-\alpha_1)\left(\frac{1}{s}\sum_{j=0}^{s} x\left[\left\lfloor\frac{C^i}{4}\right\rfloor \times 4 + j, H^i, W^i\right]\right), \tag{7}$$

where $x_i$ is the $i$-th input at location $C^i, H^i, W^i$, with $[\cdot,\cdot,\cdot]$ representing the indices along with the channel, height and width dimensions. The connection of $\mathrm{Conv}^{\mathrm{up}}$ is constructed by concatenation function to interpolate values and align with the output dimension of the convolution layer:

$$\mathrm{QConv}^{\mathrm{up}}(x) = \alpha_2 Q(\mathrm{Conv}^{\mathrm{up}}(x)) + (1-\alpha_2)\mathrm{repeat}_C(x, s), \tag{8}$$

where $\mathrm{repeat}_C(\cdot,\cdot)$ means copied along channel dimension for input $x$ by $s$ times. As for $\mathrm{Conv}^{eq}$, the connection is

$$\mathrm{QConv}^{\mathrm{eq}}(x) = \alpha_3 Q(\mathrm{Conv}^{\mathrm{eq}}(x)) + (1-\alpha_3)\mathrm{Avgpool}(x). \tag{9}$$

$\alpha_1, \alpha_2, \alpha_3$ in above equations are learnable scalars to balance the quantized and full-precision passes.

We showcase feature visualization heatmaps to illustrate the differences in features captured by the two paths and

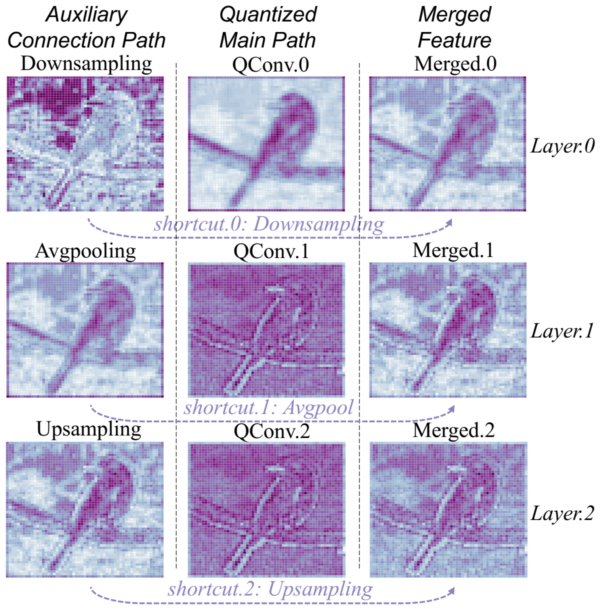


Figure 6: Heatmaps of activations in one block for a fake image of the "**bird**" by using the proposed CQB.

demonstrate the effectiveness of combining learnable quantized blocks with full-precision shortcuts for deepfake detection. And the results can be found in Sec. 4.2.

### 3.5 Shifted Logarithmic Redistribution for Texture Unfolding

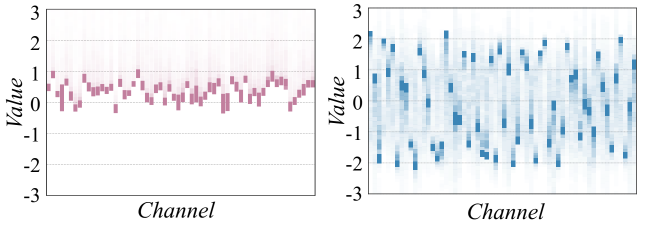
When quantizing the activations below 4-bit, we observe a significant accuracy drop due to the feature distortion caused by quantization. And the results can be found in Sec. 4.2.

Therefore, we propose a **Shifted Logarithmic Redistribution (SLR)** Quantizer for the unbalanced activation distribution, which unfolds the clustered activations near zero, assigning fine-grained quantization intervals using a non-uniform quantizer to better quantize the majority of values.

Ideally, a well-trained quantizer prefers to have the values equally distributed within the quantized grids to maximize the information entropy. However, if we use uniform quantizer to unbalanced data in Figure 7(a), there will be nearly 80% integers dealing with less than 5% of activations, while the rest 95% activations are coarsely fallen into few intervals. This inspires us to adopt non-uniform quantizer to preserve more information at post-activation function quantization, but the initialization and training of non-uniform quantizers are non-trivial. Therefore, we first apply an offset off for the input activation $X$ to shift the range into the domain of the logarithmic function:

$$\text{off} = \min(\max(|X|), \epsilon) + \tau, \tag{10}$$

where $\min(\cdot), \max(\cdot), |\cdot|$, are the minimum, maximum and absolute functions, $\epsilon \geq 1$ is a minimal offset to avoid steeper slopes of the logarithmic function in the range of $(0, 1]$ (refer to the illustration on the right part of Figure 5). And $\tau$ is a layer-wise parameter, which is trained together with the whole network, making the quantizer flexible between layers.



(a) Post-Activation Distribution

(b) Logarithmic Post-Activation Distribution

Figure 7: Distributions of post-activation function, where (a) is the original activation, (b) is the logarithmic of shifted activation.

Next, the SLR quantizer can be written as

$$Q(X) = \sum_{i=0}^{N-1} \mathbb{I}_i [\![ ln(X + \text{off}) \geq t_i ]\!], \tag{11}$$

where $ln(\cdot)$ means logarithmic function. By applying the offset, we leverage the gentler region of the logarithmic function to redistribute unbalanced parameters. Notably, the proposed SLR quantizer can be seamlessly integrated into a quantizer based on a lookup strategy. For Eq. (3), the threshold set $T$ transformed into $T' = \{e^{t_0} - \text{off}, \ldots, e^{t_{N-1}} - \text{off}\}$. By merging the offset and logarithmic operations into the set $T$, SLR quantizer introduces no extra computational overhead.

## 4 Experiments

### 4.1 Experimental Settings

**Baselines**: CNNDetection [Wang and others, 2020], Frank [Frank and others, 2020], Durall[Durall and others, 2020], Patchfor [Chai and others, 2020], F3Net [Qian and others, 2020], SelfBland [Shiohara and others, 2022], GANDetection [Mandelli and others, 2022], BiHPF [Jeong and others, 2022a], FrePGAN [Jeong and others, 2022c], LGrad [Tan *et al.*, 2023], Ojha [Ojha *et al.*, 2023] and FreqNet [Tan *et al.*, 2024a]. **Evaluation Metrics**: Following common practices [Jeong and others, 2022a; Jeong and others, 2022c; Ojha *et al.*, 2023], we use average precision (AP) and accuracy (Acc) as the evaluation metrics.

### 4.2 Preliminary experiments

**Feature Visualization.** We visualize the features for each quantized convolution layer as the *Main Path* and their corresponding full-precision connections as the *Auxiliary Path*, as shown in Figure 6. Each row has three images: the output of the shortcut connection and the quantized convolution, as well as the weighted combination of the two features. It can be observed that quantization captures the primary features, especially edge information. However, due to the rounding operations , many important information and details are lost. For instance, the background and the main subject are shown as the same color. Fortunately, the auxiliary path preserves much of the original information, particularly textures and key elements in the image (such as the bird's beak). By combining the features from both paths, the final feature map retains the outline of subject, the distinction to the background, and critical details and texture, achieving more robust and comprehensive representations.

| Methods | ProGAN | | StyleGAN | | StyleGAN2 | | BigGAN | | CycleGAN | | StarGAN | | GauGAN | | Deepfake | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | AP | Acc | AP | Acc | AP | Acc | AP | Acc | AP | Acc | AP | Acc | AP | Acc | AP | Acc | AP |
| CNNDetection | 91.4 | 99.4 | 63.8 | 91.4 | 76.4 | 97.5 | 52.9 | 73.3 | 72.7 | 88.6 | 63.8 | 90.8 | 63.9 | 92.2 | 51.7 | 62.3 | 67.1 | 86.9 |
| Frank *et al.* | 90.3 | 85.2 | 74.5 | 72.0 | 73.1 | 71.4 | 88.7 | 86.0 | 75.5 | 71.2 | 99.5 | 99.5 | 69.2 | 77.4 | 60.7 | 49.1 | 78.9 | 76.5 |
| Durall *et al.* | 81.1 | 74.4 | 54.4 | 52.6 | 66.8 | 62.0 | 60.1 | 56.3 | 69.0 | 64.0 | 98.1 | 98.1 | 61.9 | 57.4 | 50.2 | 50.0 | 67.7 | 64.4 |
| Patchfor | 97.8 | 100.0 | 82.6 | 93.1 | 83.6 | 98.5 | 64.7 | 69.5 | 74.5 | 87.2 | 100.0 | 100.0 | 57.2 | 55.4 | 85.0 | 93.2 | 80.7 | 87.1 |
| F3Net | 99.4 | 100.0 | 92.6 | 99.7 | 88.0 | 99.8 | 65.3 | 69.9 | 76.4 | 84.3 | 100.0 | 100.0 | 58.1 | 56.7 | 63.5 | 78.8 | 80.4 | 86.2 |
| SelfBland | 58.8 | 65.2 | 50.1 | 47.7 | 48.6 | 47.4 | 51.1 | 51.9 | 59.2 | 65.3 | 74.5 | 89.2 | 59.2 | 65.5 | 93.8 | 99.3 | 61.9 | 66.4 |
| GANDetection | 82.7 | 95.1 | 74.4 | 92.9 | 69.9 | 87.9 | 76.3 | 89.9 | 85.2 | 95.5 | 68.8 | 99.7 | 61.4 | 75.8 | 60.0 | 83.9 | 72.3 | 90.1 |
| BiHPF | 90.7 | 86.2 | 76.9 | 75.1 | 76.2 | 74.7 | 84.9 | 81.7 | 81.9 | 78.9 | 94.4 | 94.4 | 69.5 | 78.1 | 54.4 | 54.6 | 78.6 | 77.9 |
| FrePGAN | 99.0 | 99.9 | 80.7 | 89.6 | 84.1 | 98.6 | 69.2 | 71.1 | 71.1 | 74.4 | 99.9 | 100.0 | 60.3 | 71.7 | 70.9 | 91.9 | 79.4 | 87.2 |
| LGrad | 99.9 | 100.0 | 94.8 | 99.9 | 96.0 | 99.9 | 82.9 | 90.7 | 85.3 | 94.0 | 99.6 | 100.0 | 72.4 | 79.3 | 58.0 | 67.9 | 86.1 | 91.5 |
| Ojha *et al.* | 99.7 | 100.0 | 89.0 | 98.7 | 83.9 | 98.4 | 90.5 | 99.1 | 87.9 | 99.8 | 91.4 | 100.0 | 89.9 | 100.0 | 80.2 | 90.2 | 89.1 | **98.3** |
| FrqeNet | 99.6 | 100.0 | 90.2 | 99.7 | 88.0 | 99.5 | 90.5 | 96.0 | 95.8 | 99.6 | 85.7 | 99.8 | 93.4 | 98.6 | 88.9 | 94.4 | **91.5** | **98.5** |
| Lanzino *et al.* (1-bit) | 73.2 | 80.8 | 60.3 | 64.4 | 53.9 | 56.8 | 56.5 | 57.0 | 57.9 | 56.5 | 62.2 | 71.5 | 61.1 | 61.8 | 54.0 | 54.3 | 59.8 | 62.9 |
| MinMax (4-bit) | 91.0 | 97.5 | 74.4 | 84.4 | 73.7 | 88.7 | 48.9 | 49.0 | 50.9 | 51.0 | 78.3 | 97.8 | 47.4 | 45.7 | 54.8 | 54.2 | 64.9 | 71.1 |
| LSQ (4-bit) | 99.6 | 100.0 | 82.2 | 90.4 | 79.6 | 94.3 | 62.7 | 65.7 | 68.6 | 75.3 | 100.0 | 100.0 | 41.0 | 42.5 | 84.9 | 93.1 | 77.3 | 82.6 |
| N2UQ (3-bit) | 99.5 | 100.0 | 83.1 | 89.9 | 83.7 | 97.1 | 64.6 | 66.4 | 71.4 | 82.0 | 100.0 | 100.0 | 37.9 | 39.8 | 77.2 | 89.0 | 77.1 | 83.1 |
| Ours (3-bit) | 99.3 | 100.0 | 96.4 | 99.4 | 96.9 | 99.5 | 93.1 | 96.4 | 88.2 | 96.0 | 97.9 | 99.9 | 92.1 | 94.6 | 67.0 | 64.9 | 91.4 | 93.9 |
| Ours (2-bit) | 99.8 | 100.0 | 96.1 | 99.5 | 99.4 | 100.0 | 93.3 | 97.2 | 86.7 | 90.3 | 93.4 | 99.8 | 89.3 | 93.0 | 76.9 | 81.2 | **91.8** | 95.1 |

Table 2: Cross-GAN-Sources Evaluation on the ForenSynths. Separated by a horizontal line, the upper section presents the full-precision models and algorithms, while the lower section displays the low-bit quantized models with various quantization methods and bitwidth.

| Method | DALLE | | Glide_100_10 | | Glide_100_27 | | Glide_50_27 | | ADM | | LDM_100 | | LDM_200 | | LDM_200_cfg | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | AP | Acc | AP | Acc | AP | Acc | AP | Acc | AP | Acc | AP | Acc | AP | Acc | AP | Acc | AP |
| CNNDetection | 51.8 | 61.3 | 53.3 | 72.9 | 53.0 | 71.3 | 54.2 | 76.0 | 54.9 | 66.6 | 51.9 | 63.7 | 52.0 | 64.5 | 51.6 | 63.1 | 52.8 | 67.4 |
| Frank | 57.0 | 62.5 | 53.6 | 44.3 | 50.4 | 40.8 | 52.0 | 42.3 | 53.4 | 52.5 | 56.6 | 51.3 | 56.4 | 50.9 | 56.5 | 52.1 | 54.5 | 49.6 |
| Durall | 55.9 | 58.0 | 54.9 | 52.3 | 48.9 | 46.9 | 51.7 | 49.9 | 40.6 | 42.3 | 62.0 | 62.6 | 61.7 | 61.7 | 58.4 | 58.5 | 54.3 | 54.0 |
| Patchfor | 79.8 | 99.1 | 87.3 | 99.7 | 82.8 | 99.1 | 84.9 | 98.8 | 74.2 | 81.4 | 95.8 | 99.8 | 95.6 | 99.9 | 94.0 | 99.8 | 86.8 | 97.2 |
| F3Net | 71.6 | 79.9 | 88.3 | 95.4 | 87.0 | 94.5 | 88.5 | 95.4 | 69.2 | 70.8 | 74.1 | 84.0 | 73.4 | 83.3 | 80.7 | 89.1 | 79.1 | 86.5 |
| SelfBland | 52.4 | 51.6 | 58.8 | 63.2 | 59.4 | 64.1 | 64.2 | 68.3 | 58.3 | 63.4 | 53.0 | 54.0 | 52.6 | 51.9 | 51.9 | 52.6 | 56.3 | 58.7 |
| GANDetection | 67.2 | 83.0 | 51.2 | 52.6 | 51.1 | 51.9 | 51.7 | 53.5 | 49.6 | 49.0 | 54.7 | 65.8 | 54.9 | 65.9 | 53.8 | 58.9 | 54.3 | 60.1 |
| LGrad | 88.5 | 97.3 | 89.4 | 94.9 | 87.4 | 93.2 | 90.7 | 95.1 | 86.6 | 100.0 | 94.8 | 99.2 | 94.2 | 99.1 | 95.9 | 99.2 | 90.9 | 97.2 |
| Ojha | 89.5 | 96.8 | 90.1 | 97.0 | 90.7 | 97.2 | 91.1 | 97.4 | 75.7 | 85.1 | 90.5 | 97.0 | 90.2 | 97.1 | 77.3 | 88.6 | 86.9 | 94.5 |
| FrqeNet | 97.2 | 99.7 | 87.9 | 96.0 | 84.4 | 95.6 | 86.6 | 95.8 | 67.2 | 75.4 | 97.8 | 99.9 | 97.4 | 99.9 | 97.2 | 99.9 | 89.5 | 95.3 |
| Lanzino *et al.* (1-bit) | 55.2 | 57.9 | 59.7 | 62.5 | 59.9 | 63.2 | 61.6 | 65.0 | 57.2 | 59.3 | 60.8 | 63.5 | 61.5 | 65.4 | 57.1 | 59.6 | 59.1 | 62.0 |
| MinMax (4-bit) | 54.8 | 58.7 | 60.0 | 65.6 | 59.7 | 66.4 | 63.3 | 70.5 | 58.9 | 61.3 | 79.7 | 90.0 | 80.2 | 90.4 | 77.7 | 87.0 | 66.8 | 73.8 |
| LSQ (4-bit) | 66.0 | 94.0 | 91.9 | 99.5 | 88.2 | 99.1 | 90.3 | 99.3 | 72.1 | 79.4 | 95.2 | 99.8 | 96.0 | 99.8 | 94.0 | 99.7 | 86.7 | 96.3 |
| N2UQ (3-bit) | 71.9 | 84.5 | 87.8 | 95.6 | 83.7 | 92.8 | 85.5 | 93.8 | 68.5 | 76.7 | 91.0 | 97.3 | 90.8 | 97.0 | 90.3 | 97.1 | 83.7 | 91.8 |
| Ours (3-bit) | 82.7 | 96.0 | 96.4 | 99.3 | 94.3 | 98.9 | 93.5 | 98.8 | 86.0 | 94.6 | 97.9 | 99.5 | 97.5 | 99.4 | 97.8 | 99.5 | **93.3** | **98.2** |
| Ours (2-bit) | 82.5 | 94.0 | 96.5 | 99.3 | 96.2 | 99.5 | 95.6 | 99.4 | 82.5 | 94.0 | 95.7 | 99.3 | 94.7 | 99.2 | 95.1 | 99.3 | **92.4** | **98.4** |

Table 3: Cross-Diffusion-Sources Evaluation on the UniversalFakeDetect. Separated by a horizontal line, the upper section presents the full-precision models and algorithms, while the lower displays the low-bit quantized models with various quantization methods and bitwidth.

**Activation Visualization.** We visualize the intermediate activations at post-activation function in Figure 7(a). Each individual channel of activation are represents by x-axis, and the y-axis indicates the value distribution, with darker colors indicating higher densities. It is evident that after passing through the activation function, a large proportion of activations concentrate near zero, with only a small fraction of values much greater than zero, resulting in an imbalanced distribution. After quantization, there is always only very few effective bits represent the majority of values, leading to substantial information loss. While in Figure 7(b), after adding an offset to the parameters and mapping them to the logarithmic space, the activations are flattened and balanced, creating a more uniform distribution, friendly for quantization.

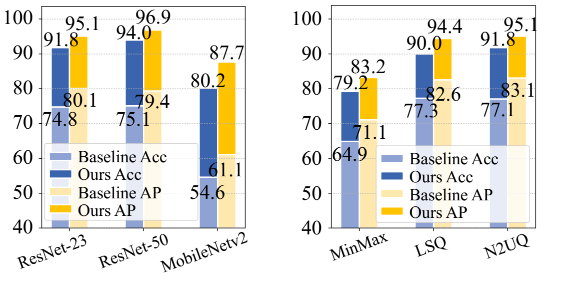
## 4.3 Detection Performance

In this section, we show the accuracy performance of our proposed lightweight designs compared to existing deepfake detection models and algorithms in Table 2 for Cross-GAN-Sources Evaluation on ForenSynths [Wang and others, 2020] and Table 3 for Cross-Diffusion-Sources Evaluation on UniversalFakeDetect [Ojha *et al.*, 2023]. Among these, only ProGAN is tested on seen data, while others are unseen.

When on the seen data domain, i.e., images generated by ProGAN, the detection accuracy of the quantized model with proposed methods is comparable to that of the models without quantization. For example, in Table 2, when we apply 2/3-bit weight and activations, our proposed quantization models have nearly no accuracy loss. And for unseen data domain, our method shows its effectiveness across different GAN and Diffusion. Under Cross-GAN-Source Evaluation, our method performs best in accuracy which is 91.8% on average, and the 95.1% AP also surpasses most of the previous methods. Especially, our model has 2.7% and improvement in Acc compared to CLIP model, while only utilizing 5% parameters and much fewer FLOPs. Meanwhile, under Cross-Diffusion-Source Evaluation, our method also achieves the best performance, which are 2.4% and 3.1% higher than the previous SOTA in Acc and AP on average performance.

## 4.4 Generalization Performance

To show the generalization ability on different backbones of our methods, we conduct experiments on different sizes of ResNets, i.e., ResNet-23/50 and MobileNet, which has a different fundamental architecture compared to residual blocks. We show the accuracy performance of the quantized models



(a) On different backbones (b) On different quantizers

Figure 8: Average performance on eight GAN-based methods.

| Method | Backbone | FLOPs (G) | Storage (MB) |
|---|---|---|---|
| Frank *et al.* | 4-Layer CNN | 0.0238 | 0.68 |
| FreqNet | CNN | 5.1 | 1.9 |
| F3Net | Xception | 8.4 | 22.9 |
| CNNDetection | ResNet50 | 4.8 | 25.6 |
| Ojha *et al.* | CLIP_ResNet50 | 12.2 | 25.6 |
| Ojha *et al.* | CLIP_VIT-L/14 | 162 | 88.3 |
| SelfBland | ResNet50+ | 8.34 | 44.6 |
| LGrad | Discriminator+ | 75 | 46.6 |
| GANDetection | EfficentNetB4×5 | 21 | 95 |
| Lanzino *et al.* | BNext-T [1-bit] | 0.89 | 13.3 |
| Lanzino *et al.* | BNext-M [1-bit] | 3.4 | 46.5 |
| FP Basline | ResNet-23 | 2.7 | 62 |
| LSQ | ResNet-23 [4-bit] | 0.37 | 7.9 |
| N2UQ | ResNet-23 [3-bit] | 0.30 | 6.9 |
| Ours | ResNet-23 [3-bit] | 0.32 | 6.9 |
| Ours | ResNet-23 [2-bit] | 0.25 | 5.0 |

Table 4: Efficiency of methods, backbones and quantized bitwidth. For low-bit quantized methods, we calculate OPs as theoretical computation complexity, while for full-precision models, we use FLOPs.

using naive quantization techniques (light blue/yellow bars), or equipped with our proposed methods (dark blue/yellow bars) in Figure 8(a). All the cases are tested on ForenSynths, and we only show the average Acc and AP for simplicity. The experimental results demonstrate that our proposed method consistently and significantly improves detection performance across both ResNet and MobileNet architectures, achieving more than a 20% improvement in Acc and a 15% improvement in AP. Notably, due to the inherently lightweight structure of MobileNet, directly applying conventional quantization methods results in severe accuracy degradation. In contrast, our method achieves remarkable improvements, boosting Acc by 25.6% and AP by 26.6%. These gains are mainly attributed to our CQB module, which enhances the information density of the MobileNet architecture.

Regarding different quantizers, we evaluated our quantization method with MinMax,LSQ [Esser *et al.*, 2020], and N2UQ [Liu *et al.*, 2022] quantizers. As shown in Figure 8(b), our SLR consistently achieves stable improvements across all quantizers, which demonstrates that by leveraging shift and logarithmic redistribution, SLR effectively mitigates the issue of unbalanced parameter distribution, thereby significantly enhancing the performance.

### 4.5 Efficiency Performance

We analyze the efficiency of various methods and backbones in Table 4, reporting theoretical GFLOPs and storage (MB).

The methods can be roughly categorized into two kinds depending on the data preprocessing. Methods like [Durall and others, 2020; Frank and others, 2020; Tan *et al.*, 2024a] are based on frequency analysis, which may have relatively fewer model storage but increased FLOPs due to DCT computation. Other approaches use diverse backbones such as CLIP-ViT, ResNet, or EfficientNet. CLIP-based models are the most resource-intensive, with up to 162 GFLOPs and 88.3MB storage. With quantization, the FLOPs can be reduced by higher throughput computations, and the storage can also be compact by packing the low-bit parameters together. For instance, BNext-T [Lanzino *et al.*, 2024] uses binarization to achieve 0.89 GFLOPs and 13.3MB, though at a cost of significant accuracy loss. Therefore, we adopt 2/3-bit quantization in our method, allowing us to use smaller models while maintaining the accuracy. Based on ResNet-23, our model achieves strong compression, requiring only 0.25 GFLOPs and 5.0MB under 2-bit, while maintaining competitive accuracy.

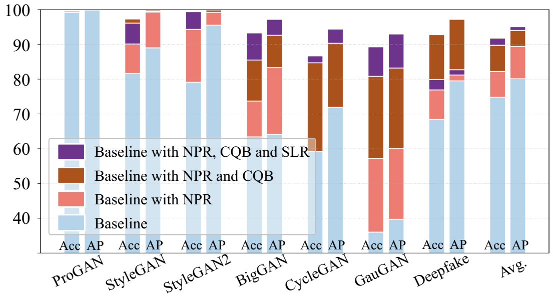


Figure 9: Ablation study on Cross-GAN-Sources Evaluations.

### 4.6 Ablation Study

We evaluate each proposed component in Figure 9 on Cross-GAN-Source Evaluations. The baseline (blue bar) uses 2-bit N2UQ on RGB images. Replacing RGB with NPR (pink bar) significantly boosts performance by 7.4% in Acc and 9.3% in AP, highlighting NPR is more quantization-friendly due to its compact value range and finer quantization intervals.

Adding shortcut connections in CQB (brown bar) yields a further 7.5% Acc gain by preserving fine-grained texture through the full-precision auxiliary path. Finally, our SLR quantizer (purple bar), designed for post-activation imbalance, brings a 2.1% improvement. These results confirm that combining NPR, CQB, and SLR effectively enhances detection performance under low-bit quantizationacross different generation models.

## 5 Conclusion

In this paper, we focus on the lightweight deepfake detection models and propose low-bit quantization method. It includes a Connected Quantized Block (CQB), which preserves fine details and textures to compensate for the forgery features in quantized convolutional layers, and the Shifted Logarithmic Redistribution (SLR) Quantizer, which unfolds unbalanced activations to make them more quantization-friendly and reduce the information loss. The proposed quantized deepfake detection model achieves a 10.8× reduction in computation and a 12.4× decrease in model storage, while maintaining accuracy comparable to full-precision or larger models.

## Acknowledgments

This work was supported by Beijing NSF (No. L242021), National NSF of China (No. U24B20179), the Talent Fund of BJTU (No. 2024XKRC047) and the Open Project of Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University (No. MMC202406).

## References

[Asnani *et al.*, 2023] Vishal Asnani, Xi Yin, Tal Hassner, and Xiaoming Liu. Reverse engineering of generative models: Inferring model hyperparameters from generated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Cao and others, 2022] Junyi Cao et al. End-to-end reconstruction-classification learning for face forgery detection. In *CVPR*, pages 4113–4122, 2022.

[Chai and others, 2020] Lucy Chai et al. What makes fake images detectable? understanding properties that generalize. In *ECCV*, pages 103–120. Springer, 2020.

[Chen and others, 2022] Liang Chen et al. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *CVPR*, pages 18710–18719, 2022.

[Ding *et al.*, 2023] Hongwei Ding, Yu Sun, Nana Huang, Zhidong Shen, and Xiaohui Cui. Tmg-gan: Generative adversarial networks-based imbalanced learning for network intrusion detection. *IEEE Transactions on Information Forensics and Security*, 19:1156–1167, 2023.

[Ding *et al.*, 2024] Yifu Ding, Weilun Feng, Chuyan Chen, Jinyang Guo, and Xianglong Liu. Reg-ptq: Regression-specialized post-training quantization for fully quantized object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16174–16184, 2024.

[Durall and others, 2020] Ricard Durall et al. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *CVPR*, pages 7890–7899, 2020.

[Esser *et al.*, 2019] Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. *arXiv preprint arXiv:1902.08153*, 2019.

[Esser *et al.*, 2020] Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. Learned step size quantization. In *International Conference on Learning Representations*, 2020.

[Frank and others, 2020] Joel Frank et al. Leveraging frequency analysis for deep fake image recognition. In *ICML*, pages 3247–3258. PMLR, 2020.

[He and others, 2021] Yang He et al. Beyond the spectrum: Detecting deepfakes via re-synthesis. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2534–2541. International Joint Conferences on Artificial Intelligence Organization, 2021.

[Hernández *et al.*, 2024] Nicolás Hernández, Francisco Almeida, and Vicente Blanco. Optimizing convolutional neural networks for iot devices: performance and energy efficiency of quantization techniques. *The Journal of Supercomputing*, pages 1–20, 2024.

[Jeong and others, 2022a] Yonghyun Jeong et al. Bihpf: Bilateral high-pass filters for robust deepfake detection. In *WACV*, pages 48–57, 2022.

[Jeong and others, 2022b] Yonghyun Jeong et al. Fingerprintnet: Synthesized fingerprints for generated image detection. In *ECCV*, pages 76–94. Springer, 2022.

[Jeong and others, 2022c] Yonghyun Jeong et al. Frepgan: robust deepfake detection using frequency-level perturbations. In *AAAI*, volume 36, pages 1060–1068, 2022.

[Ju and others, 2022] Yan Ju et al. Fusing global and local features for generalized ai-synthesized image detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3465–3469. IEEE, 2022.

[Jung *et al.*, 2019] Sangil Jung, Changyong Son, Seohyung Lee, Jinwoo Son, Jae-Joon Han, Youngjun Kwak, Sung Ju Hwang, and Changkyu Choi. Learning to quantize deep networks by optimizing quantization intervals with task loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4350–4359, 2019.

[Lanzino *et al.*, 2024] Romeo Lanzino, Federico Fontana, Anxhelo Diko, Marco Raoul Marini, and Luigi Cinque. Faster than lies: Real-time deepfake detection using binary neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3771–3780, June 2024.

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

[Liu *et al.*, 2020] Zechun Liu, Zhiqiang Shen, Marios Savvides, and Kwang-Ting Cheng. Reactnet: Towards precise binary neural network with generalized activation functions. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 143–159. Springer, 2020.

[Liu *et al.*, 2022] Zechun Liu, Kwang-Ting Cheng, Dong Huang, Eric P Xing, and Zhiqiang Shen. Nonuniform-to-uniform quantization: Towards accurate quantization via generalized straight-through estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4942–4952, 2022.

[Liu *et al.*, 2023] Jun Liu, Jiantao Zhou, Haiwei Wu, Weiwei Sun, and Jinyu Tian. Generating robust adversarial examples against online social networks (osns). *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(4):1–26, 2023.

[Mandelli and others, 2022] Sara Mandelli et al. Detecting gan-generated images by orthogonal training of multiple cnns. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3091–3095. IEEE, 2022.

[Ojha *et al.*, 2023] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023.

[Qian and others, 2020] Yuyang Qian et al. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, pages 86–103. Springer, 2020.

[Shiohara and others, 2022] Kaede Shiohara et al. Detecting deepfakes with self-blended images. In *CVPR*, pages 18720–18729, 2022.

[Tan and others, 2023] Chuangchuang Tan et al. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *CVPR (CVPR)*, pages 12105–12114, June 2023.

[Tan *et al.*, 2023] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12105–12114, June 2023.

[Tan *et al.*, 2024a] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space learning, 2024.

[Tan *et al.*, 2024b] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024.

[Tan *et al.*, 2025] Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7184–7192, 2025.

[Tao *et al.*, 2021] Renshuai Tao, Yanlu Wei, Xiangjian Jiang, Hainan Li, Haotong Qin, Jiakai Wang, Yuqing Ma, Libo Zhang, and Xianglong Liu. Towards real-world x-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10923–10932, 2021.

[Tao *et al.*, 2022] Renshuai Tao, Hainan Li, Tianbo Wang, Yanlu Wei, Yifu Ding, Bowei Jin, Hongping Zhi, Xianglong Liu, and Aishan Liu. Exploring endogenous shift for cross-domain detection: A large-scale benchmark and perturbation suppression network. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21157–21167. IEEE, 2022.

[Tao *et al.*, 2025a] Renshuai Tao, Manyi Le, Chuangchuang Tan, Huan Liu, Haotong Qin, and Yao Zhao. Oddn: Addressing unpaired data challenges in open-world deepfake detection on online social networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 799–807, 2025.

[Tao *et al.*, 2025b] Renshuai Tao, Shijie Tang, Haotong Qin, Wei Wang, Yunchao Wei, and Yao Zhao. Lednet: A multimodal foundation model for robust deepfake detection. *Science China Information Sciences*, 68(6):160106, 2025.

[Vice *et al.*, 2024] Jordan Vice, Naveed Akhtar, Richard Hartley, and Ajmal Mian. Bagm: A backdoor attack for manipulating text-to-image generative models. *IEEE Transactions on Information Forensics and Security*, 2024.

[Wang and others, 2020] Sheng-Yu Wang et al. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, pages 8695–8704, 2020.

[Wang and others, 2021] Chengrui Wang et al. Representative forgery mining for fake face detection. In *CVPR*, pages 14923–14932, 2021.

[Wang and others, 2023] Zhendong Wang et al. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22445–22455, October 2023.

[Wang *et al.*, 2023a] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *CVPR*, pages 4129–4138, 2023.

[Wang *et al.*, 2023b] Zhendong Wang, Hui Chen, Shuxin Yang, Xiao Luo, Dahai Li, and Junling Wang. A lightweight intrusion detection method for iot based on deep learning and dynamic quantization. *PeerJ Computer Science*, 9:e1569, 2023.

[Wei *et al.*, 2020] Yanlu Wei, Renshuai Tao, Zhangjie Wu, Yuqing Ma, Libo Zhang, and Xianglong Liu. Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module. In *Proceedings of the 28th ACM international conference on multimedia*, pages 138–146, 2020.

[Wu *et al.*, 2023] Haiwei Wu, Jiantao Zhou, Xinyu Zhang, Jinyu Tian, and Weiwei Sun. Robust camera model identification over online social network shared images via multi-scenario learning. *IEEE Transactions on Information Forensics and Security*, 2023.

[Wu *et al.*, 2024] Siyang Wu, Jiakai Wang, Jiejie Zhao, Yazhe Wang, and Xianglong Liu. Napguard: Towards detecting naturalistic adversarial patches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24367–24376, 2024.

[Yin *et al.*, 2024] Zixin Yin, Jiakai Wang, Yisong Xiao, Hanqing Zhao, Tianlin Li, Wenbo Zhou, Aishan Liu, and Xianglong Liu. Improving deepfake detection generalization by invariant risk minimization. *IEEE Transactions on Multimedia*, 26:6785–6798, 2024.

[Zeng *et al.*, 2024] Kai Zeng, Kejiang Chen, Jiansong Zhang, Weiming Zhang, and Nenghai Yu. Towards secure and robust steganography for black-box generated images. *IEEE Transactions on Information Forensics and Security*, 2024.