Context-Aware Deepfake Detection for Securing AI-Driven Financial Transactions

Changcun Liu†, Guisheng Zhang†, Siyou Guo, Qilei Li, Gwanggil Jeon, Mingliang Gao*

***Abstract*—The rapid advancement of deepfake technology has threatened the community's sense of security, particularly in the context of face-based payment systems. Thus, deepfake detection has emerged as a critical issue demanding immediate attention. However, the generalization performance of existing detection models is limited as they are overly reliant on specific forged features while ignoring the common forged features. To address this problem, we introduce the Context-Aware Decoupling Network (CADNet) for deepfake detection. Specifically, a Context Self-Calibration (CSC) module is constructed to guide the network to focus on local forged regions. It enlarges possible regions to increase the likelihood of forgery cues. Meanwhile, a Frequency Domain Decoupling (FDD) module is introduced to extract and fuse different frequency components. It realizes the collaborative representation optimization of global semantics and local details. The experimental results prove that the proposed model exhibits strong generalization capability across multiple standard datasets. It achieves average AUC values of 98.64% for in-domain evaluation and 75.52% for cross-dataset generalization.**

***Index Terms*—Deepfake detection, Facial payment, Generalization, Frequency domain decoupling, Context self-calibration.**

I. Introduction

With the rapid development of the internet and the flourishing digital economy, digital finance has become a pivotal driver of economic growth. In this context, it is increasingly vital to ensure secure transactions, particularly in facial payment systems. However, with the rapid development of Artificial Intelligence Generated Content (AIGC), the misuse of forged facial images has significantly increased in facial recognition payment systems. As digital finance expands, the risks posed by fraudulent facial images in the financial sector have become increasingly acute. To address this challenge, the academic community has introduced deepfake detection methods and conducted systematic and extensive research on deepfake detection [1]–[3].

Existing deepfake detection methods typically perform well when training and testing data are generated using the same deepfake techniques [4]. However, in real-world applications, testing data may be generated using unknown methods, and thus, the generalization performance is degraded. To improve the generalization performance of deepfake detection, some researchers have made some attempts in the spatial and frequency domains. Yan *et al.* [5] improved the generalization by mining the common features in different forgery techniques. Liu *et al.* [6] introduced a spatial-phase shallow learning method to improve the generalization ability of forgery face detection. Qian *et al.* [7] developed a frequency-based face forgery detection network to enhance the generalization capability. Luo *et al.* [8] improves the generalization of forgery detection systems by exploiting high-frequency features. Although these methods aim to improve the generalization ability of detection networks, their performance remains limited. This is because deepfake detection methods overfit to specific prominent regions of the deepfake images [9].

To improve the generalization of the deepfake detection model, we propose the CADNet for deepfake detection. The CADNet mainly comprises two key modules, namely Context Self-Calibration (CSC) and Frequency Domain Decoupling (FDD). The CSC module retains the original input characteristics and extracts local features. It also prompts the network to explore more forged content by examining as many regions as possible. Additionally, the FDD module extracts and fuses the low-frequency and three high-frequency components. The low-frequency component is used to capture the global pattern, and the three directional high-frequency components are used to retain detailed spatial information. In the process of high and low frequency feature processing, the EffConv block is designed to eliminate redundant feature representations. In sum, the contributions of this work are three-fold:

- A CADNet model is proposed to enhance the generalization and accuracy of the deepfake detection.
- A CSC module is introduced to guide the network to focus on the forged regions and assess specific non-critical regions effectively.
- An FDD module is developed to extract and fuse low-frequency and high-frequency components to enhance feature extraction.

The rest of the paper is structured as follows: Section II presents the related work. Section III illustrates the proposed method in detail. Section IV analyses the experimental results. The paper is concluded in Section V.

Changcun Liu, Guisheng Zhang, Siyou Guo, and Mingliang Gao are with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, 255000, China. (Email: 24404020554@stumail.sdut.edu.cn, 22504030001@stumail.sdut.edu.cn, 23504030565@stumail.sdut.edu.cn, and mlgao@sdut.edu.cn.)

Gwanggil Jeon is with the Department of Embedded Systems Engineering, Incheon National University, Incheon, 22012, South Korea (Email:ggjeon@gmail.com).

Qilei Li is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, United Kingdom (Email: q.li@qmul.ac.uk).

* Mingliang Gao is the corresponding author.

† Changcun Liu and Guisheng Zhang contributed equally to this work.

II. Related work

A. Deepfake detection

Deepfake is a generative technique that uses methods such as Generative Adversarial Network (GAN) and Diffusion model to realistically synthesize or manipulate facial images. In digital finance and other sensitive domains, the misuse of deepfake technology can lead to significant property damage [10]. Therefore, detecting fake facial images is essential to prevent fraudulent activities and mitigate potential risks. Various approaches have been proposed for deepfake detection [11]–[16]. In the early stages, deepfake detection methods focus mainly on identifying abnormal features [17] or forged traces [18] in facial images to determine their authenticity. However, with the emergence of various new deepfake generation models, these methods exhibit poor detection performance on advanced deepfake techniques, and these results in poor generalization performance.

To enhance the generalization capability of deepfake detection methods, many deep learning techniques have been extensively adopted [19], benefiting from their strong feature modeling capabilities. Cao *et al.* [20] proposed an end-to-end reconstruction-classification learning framework, termed RECCE, to improve generalization in deepfake detection. It integrates multi-scale graph reasoning and reconstruction-guided attention mechanisms to detect forged images with unknown patterns. Ni *et al.* [21] proposed a consistent representation learning model to improve generalization in detecting unknown forgery patterns. It introduces a consistent representation learning framework that explicitly constrains feature consistency across different data augmentations using a cosine similarity loss. Dang *et al.* [22] introduced an attention-driven deepfake detection model to identify manipulated facial images and locate tampered facial regions. However, these methods tend to overly rely on specific regional features and ignore the common forged features. Thus, their generalization capabilities are limited. To this aim, we propose the CADNet through context-dependent insights and decoupled feature analysis.

B. Global context attention

The global context attention is widely used in deepfake detection for capturing global information efficiently. It bridges the gap between local and global features and enhances model adaptability and performance. Li *et al.* [23] proposed a long-distance attention mechanism that integrates global information to enhance the comprehensiveness of feature representation. Wang *et al.* [24] introduced the multi-domain attention mapping learning mechanism to enhance the adaptability of the model to diverse deepfakes. Additionally, Das *et al.* [25] developed a gated context attention mechanism to filter and aggregate relevant contextual information from coarse feature maps. In this work, we build a Context Self-Calibration (CSC) module to guide the network to focus on the forged regions and assess specific non-critical regions.

C. Frequency domain analysis

Spatial domain processing methods directly analyze the raw pixels to detect forgery. They overlook frequency domain anomalies generated during the forgery process. In contrast, the frequency domain processing methods detect forgery through spectral transformations. They focus on local frequency domain features of the image, such as high-frequency components like edges and textures, and low-frequency components like smooth regions. Thus, the frequency domain approaches are more suitable for capturing pixel-level artifacts or frequency anomalies generated during the forgery process.

Wavelet transform is a common tool to capture features by multi-resolution analysis in forgery detection. Wang *et al.* [26] utilized wavelet transform to derive frequency-domain features and improve the generalization of the deepfake detection model across datasets. Gao *et al.* [27] employed the discrete wavelet transform to enhance global high-frequency features in complex forgery detection. Liu *et al.* [28] proposed a multi-scale wavelet transformer architecture for deepfake detection. Wolter *et al.* [29] proposed a wavelet packet-based approach for deepfake detection. In this work, we propose a Frequency Domain Decoupling module to decouple and integrate low-frequency and high-frequency components. The proposed CADNet enhances the model's generalization capability. It uses the integration of multi-level high-frequency features, spatial features, and a frequency-guided attention mechanism.

III. Proposed Model

A. Overview

The schematic of the CADNet is shown in Fig. 1. The image pair is processed by two independent encoders. A face encoder extracts the initial face features, while a background encoder extracts the initial background features. The CSC module enhances the face and background features and derives F_f and F_b (or R_f and R_b) by emphasizing local features and more feature regions. Next, the initial specific features and common features are obtained by separating the facial features through convolutional layers. Subsequently, the initial specific features and common features are enhanced using the FDD module to derive F_s and F_c (R_s and R_c). The common head is designed to assess the authenticity of the images and enhance generalization to unseen forgery techniques. The specific head is a multi-class classification that distinguishes between various forgery techniques.

In addition, the reconstruction module is divided into the self-reconstruction part and the cross-reconstruction part. The self-reconstruction part is to ensure that the decoupled features are consistent with the original image, and the cross-reconstruction part is to verify the independence of the decoupled features. The background and face features are processed using three Adaptive Instance Normalization (AdaIN) layers [30], as well as four convolutional and four upsampling operations. Then, the reassembled features are processed by the decoder to reconstruct facial images. The reconstructed images are used for reconstruction loss calculation. The reconstruction loss function facilitates enhanced feature decoupling between background and facial features through image reconstruction.

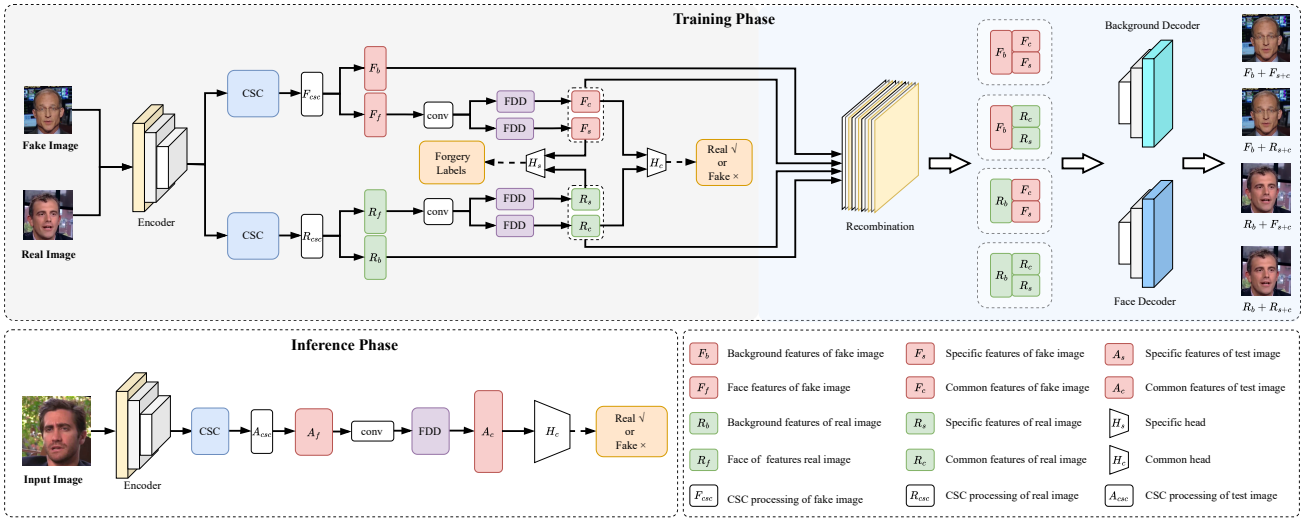

Fig. 1. The overview framework of the CADNet model for deepfake detection.

B. Context self-calibration module

To enhance the model's ability to capture local key regions and global contextual dependencies in facial images, a Context Self-Calibration (CSC) module is built. The schematic of the CSC module is shown in Fig. 2.

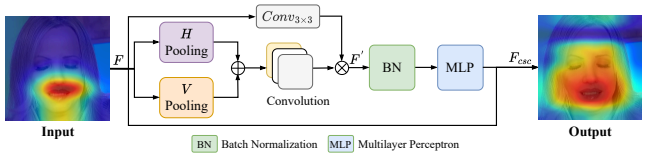

Fig. 2. The overview of the CSC module. H refers to horizontal pooling. V refers to vertical pooling.

Specifically, the residual feature information is compressed along the vertical axis using horizontal pooling. It extracts global information from each column and generates horizontal global context features. Similarly, feature information is compressed along the horizontal axis using vertical pooling. Subsequently, the horizontal and vertical pooling results are combined to generate a fused global context feature map. The fusion formula is represented as:

$$R\_attention = HP(F) \oplus VP(F), \tag{1}$$

where F is the input feature. $HP(F)$ and $VP(F)$ denote the horizontal and vertical axial vector, respectively. $\oplus$ represents the broadcast addition operation.

The map $R\_attention$ is processed through a convolutional layer. Meanwhile, a 3×3 convolution is performed on the input feature F. Then, the results of these two operations are multiplied element-wise to generate $F^{'}$. The feature $F^{'}$ can be described as:

$$F^{'} = Conv(R\_attention) \otimes Conv_{3\times3}(F), \tag{2}$$

where $Conv_{3\times3}(\cdot)$ is a 3×3 convolution. $\otimes$ represents the element-wise multiplication. After that, the feature $F^{'}$ is passed through Batch Normalization (BN) and a Multilayer Perceptron (MLP) to obtain the global features. Finally, the model adds the global features and the input features to produce the final output. The process is formulated as:

$$F_{csc} = \text{MLP}(\text{BN}(F^{'})) + F, \tag{3}$$

where F_{csc} is the output feature of the CSC module. $\text{MLP}(\cdot)$ and $\text{BN}(\cdot)$ represent the multilayer perceptron and batch normalization, respectively.

C. Frequency domain decoupling module

The schematic of the FDD module is shown in Fig. 3. Within the FDD module, the feature map undergoes decomposition through the Haar wavelet transform. This process produces a low-frequency component along with high-frequency components in horizontal, vertical, and diagonal orientations. The Haar wavelet transform formula is expressed as:

$$\begin{aligned} l_r(i,j) &= \frac{F_f(i,2j-1) + F_f(i,2j)}{2}, \\ d_r(i,j) &= F_f(2i-1,j) - F_f(2i,j), \end{aligned} \tag{4}$$

where i and j are the 2D coordinate indices, and they are employed to locate the spatial position of the feature explicitly. $l_r(i,j)$ is a low-frequency component. F_f is the input feature. $d_r(i,j)$ is a diagonal high-frequency component.

Then, the low-frequency feature and three high-frequency features can be obtained as:

$$\begin{aligned} l(i,j) &= \frac{l_r(2i-1,j) + l_r(2i,j)}{2}, \\ h_L(i,j) &= l_r(2i-1,j) - l_r(2i,j), \\ h_V(i,j) &= d_r(2i-1,j) + d_r(2i,j), \\ h_D(i,j) &= d_r(2i-1,j) - d_r(2i,j), \end{aligned} \tag{5}$$

where $l(i,j)$ denotes the initial low frequency component. $h_L(i,j)$ and $h_V(i,j)$ represent the horizontal high-frequency and vertical high-frequency components, respectively. $h_D(i,j)$

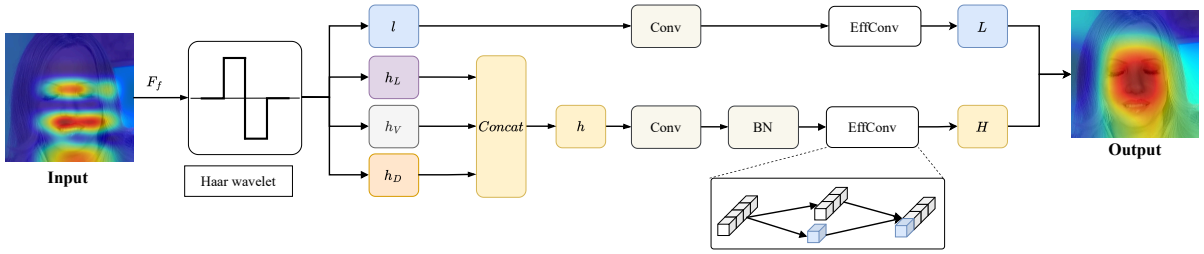

Fig. 3. The overview of the FDD module. l and h refer to the low-frequency and high-frequency features. h_L and h_V denote the horizontal high-frequency and vertical high-frequency features, respectively. h_D refers to the diagonal high-frequency feature.

is the diagonal high-frequency component. The three high-frequency features are concatenated to generate the initial high-frequency feature. Subsequently, the initial high-frequency features are refined through convolution normalization and an EffConv block to produce the final high-frequency features. The initial low-frequency features are processed through a convolutional operation followed by an EffConv block to obtain the final low-frequency representations.

The final low-frequency and high-frequency components are formulated as:

$$\begin{aligned} L &= \text{EffConv}(\text{BN}(Conv_{1\times1}(l))), \\ H &= \text{EffConv}(\text{BN}(Conv_{1\times1}(Concat(h_L, h_V, h_D)))), \end{aligned} \tag{6}$$

where L and H are the final low-frequency component and high-frequency component, respectively. EffConv$(\cdot)$ denotes the efficient convolution. $BN(\cdot)$ represents the batch normalization. $Conv_{1\times1}(\cdot)$ is the convolution kernel with a size of 1×1. $Concat(\cdot)$ represents the concatenation operation applied to three high-frequency features.

By employing the FDD module, the proposed method maintains high-frequency detail representation while simultaneously integrating low-frequency global structural information.

D. Loss function

The loss function is a combination of classification loss, contrastive regularization loss, and reconstruction loss.

1) Classification loss: The classification loss consists of two components: the binary classification loss and the Specific forgery classification loss. **Binary classification loss** ($\mathcal{L}_{\text{ce}}^{c}$): It supervises the model to learn common forgery features across different methods. The loss $\mathcal{L}_{\text{ce}}^{c}$ is denoted as:

$$\mathcal{L}_{\text{ce}}^{c} = \mathcal{L}_{\text{ce}}\left(\mathcal{H}_c\left(A_c\right), y_i\right), \tag{7}$$

where $\mathcal{H}_c(\cdot)$ is the head for common forgery features. A_c is the common fingerprint, and $y_i \in \{\text{fake}, \text{real}\}$. **Specific forgery classification loss** ($\mathcal{L}_{\text{ce}}^{s}$): It is employed to identify the specific forgery method. The loss is formulated as:

$$\mathcal{L}_{\text{ce}}^{s} = \mathcal{L}_{\text{ce}}\left(\mathcal{H}_s\left(A_s\right), y_i'\right), \tag{8}$$

where $\mathcal{H}_s(\cdot)$ is the head for specific forgery features. A_s is the specific fingerprint, and $y_i' \in \{\text{fake}, \text{GAN}_1, \text{GAN}_2, \dots\}$.

2) Contrastive regularization loss: The contrastive regularization loss $\mathcal{L}_{\text{con}}$ is formulated as:

$$\mathcal{L}_{\text{con}} = \max\left(\|x_A - x_P\|_2 - \|x_A - x_N\|_2 + \alpha, 0\right), \tag{9}$$

where α is a margin hyperparameter. x_A is the anchor. x_N is the dissimilar counterpart. x_P is the dissimilar counterpart. It ensures a generalizable representation by distinguishing real and fake images and method-specific forgeries.

3) Reconstruction loss: The reconstruction loss includes self-reconstruction loss and cross-reconstruction loss. The self-reconstruction loss ensures that the reconstructed image aligns with the original by combining the background and face from the same source. The cross-reconstruction loss verifies feature independence by combining a real background with a forged face and vice versa. The final reconstruction loss is expressed as:

$$\begin{aligned} \mathcal{L}_{\text{rec}}^{s} &= \|F - D(F_b + F_{s+c})\|_1 + \|R - D(R_b + R_{s+c})\|_1, \\ \mathcal{L}_{\text{rec}}^{c} &= \|F - D(F_b + R_{s+c})\|_1 + \|R - D(R_b + R_{s+c})\|_1, \\ \mathcal{L}_{\text{rec}} &= \mathcal{L}_{\text{rec}}^{s} + \mathcal{L}_{\text{rec}}^{c}, \end{aligned} \tag{10}$$

where F and R represent the fake and real images. $D(F_b + F_{s+c})$ and $D(R_b + R_{s+c})$ represent the self-reconstructed images processed by the background decoder and the face decoder. $D(F_b + R_{s+c})$ and $D(R_b + F_{s+c})$ represent the cross-reconstructed images processed by the background decoder and the face decoder. $\mathcal{L}_{\text{rec}}^{s}$, $\mathcal{L}_{\text{rec}}^{c}$, and $\mathcal{L}_{\text{rec}}$ correspond to self-reconstruction loss, cross-reconstruction loss, and final reconstruction loss.

4) Final loss: The final loss is to balance classification, contrastive, and reconstruction objectives to improve feature disentanglement and detection accuracy. It is obtained as:

$$\mathcal{L} = \mathcal{L}_{\text{ce}}^{c} + \lambda_1 \mathcal{L}_{\text{ce}}^{s} + \lambda_2 \mathcal{L}_{\text{con}} + \lambda_3 \mathcal{L}_{\text{rec}}, \tag{11}$$

where λ_1, λ_2, and λ_3 are weights for specific loss components and their values strictly follow the configuration established in the baseline Uncovering Common Features (UCF) [5].

IV. Experimental Results and Analysis

A. Datasets

Experiments are conducted on five deepfake datasets, namely FaceForensics++ (FF++) [31], Celeb-DF-v1 [32], Celeb-DF-v2 [32], DeepFake Detection Challenge Preview (DFDCP) [33] and FaceShifter [34]. The pre-processing for these datasets follows the criteria in work [35].

TABLE I
COMPARATIVE RESULTS ON INTRA-DATASET VALIDATION. THE TRAINING DATASET IS THE FF++ DATASET, AND THE METRIC IS AUC. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Methods	FF++	DF	F2F	FS	NT	Avg.
Meso4 [36]	0.6077	0.6771	0.6170	0.5946	0.5701	0.6133
MesoIncep [36]	0.7583	0.8542	0.8087	0.7421	0.6517	0.7630
CNN-Aug [37]	0.8493	0.9048	0.8788	0.9026	0.7313	0.8534
Xception [31]	0.9637	0.9799	0.9785	0.9833	0.9385	0.9688
EfficientB4 [38]	0.9567	0.9757	0.9758	0.9797	0.9308	0.9637
Capsule [39]	0.8421	0.8669	0.8634	0.8734	0.7804	0.8452
FWA [40]	0.8765	0.9210	0.9000	0.8843	0.8120	0.8788
X-ray [41]	0.9592	0.9794	0.9872	0.9871	0.9290	0.9684
FFD [22]	0.9624	0.9803	0.9784	0.9853	0.9306	0.9674
CORE [21]	0.9638	0.9787	0.9803	0.9823	0.9339	0.9678
Recce [20]	0.9621	0.9797	0.9779	0.9785	0.9357	0.9668
UCF [5]	0.9705	0.9883	0.9840	0.9896	0.9441	0.9753
F3Net [7]	0.9635	0.9793	0.9796	0.9844	0.9354	0.9684
SPSL [6]	0.9610	0.9781	0.9754	0.9829	0.9299	0.9655
SRM [8]	0.9576	0.9733	0.9696	0.9740	0.9295	0.9609
CADNet (Ours)	**0.9854**	**0.9940**	**0.9929**	**0.9933**	**0.9663**	**0.9864**

The FF++ dataset is a comprehensive dataset for training and evaluating deep learning models in facial manipulation detection. With over 500,000 forged images, it surpasses similar datasets and aids in developing robust detection models. The dataset includes four major forgery techniques: DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT), which reflect the diverse and advanced trends in facial forgery technology.

The Celeb-DF-v1 dataset is a foundational deepfake dataset designed to advance research in detecting manipulated media content. It contains 795 real videos featuring 59 celebrity subjects and 5,639 high-quality deepfake videos. The dataset incorporates early deepfake generation techniques, such as neural network-driven facial synthesis and blending.

The Celeb-DF-v2 dataset does not contain artifacts like unnatural lip movements and geometric distortions. It provides high-resolution videos with an equal mix of real and fake content generated using advanced techniques, which increases the challenge for detection models. As a benchmark dataset, it plays a key role in advancing deepfake detection methods.

The DFDCP dataset provides high-quality real and forged videos that use diverse facial manipulation techniques. It emphasizes realistic scenarios with variations in lighting, angles, and resolutions. This makes it ideal for training and evaluating detection models and supports the development of robust deepfake detection systems.

The FaceShifter dataset is an advanced face-swapping dataset containing highly realistic facial forgeries with minimal artifacts. Its two-stage architecture preserves facial identity and ensures consistency in lighting and background. Using a feature-based generator and refinement network, it delivers seamless results and sets a new benchmark in deepfake detection research.

B. Evaluation metrics

The proposed network is evaluated using four standard binary classification metrics, namely Area Under the Curve (AUC), Accuracy (ACC), Average Precision (AP), and Equal Error Rate (EER). The AUC quantifies the ability of the model to differentiate classes and indicates the probability of correct classification. The ACC represents the model's overall performance in the sample and measures the proportion of samples correctly classified by the model. The AP evaluates the balance between the accuracy and the recall rate of the model at different thresholds. The EER represents the threshold at which the false acceptance rate and false rejection rate converge.

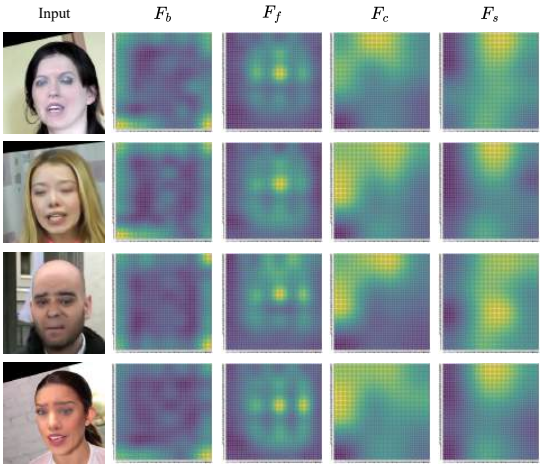

Fig. 4. Visualizations of extracted intermediate feature maps.

C. Implementation details

In this work, we used the Adam [42] optimization algorithm with a learning rate of 2×10^{-4}. The batch size was set to 32. For the Xception [31] detector, we took their official model and initialized the parameters with pre-training on ImageNet. The margin α in Eq. (9) is set to 3. In the final loss function in Eq. (11), the λ_1, λ_2, and λ_3 were set as 0.1, 0.05, and 0.3. All experiments were implemented in PyTorch [43] with an NVIDIA GeForce RTX 3090 Ti GPU.

D. Experimental results

To maintain strict experimental controls, neither data augmentation techniques nor external training data were incorporated. The model was exclusively trained on the FF++ dataset without additional synthetic or manipulated samples.

1) Intra-dataset validation: To conduct intra-testing, all methods are trained on the FF++ dataset and evaluated on both the FF++ dataset and its subsets. The comparative results between different methods are presented in Table I. Compared to other methods, the proposed method achieves superior performance on the FF++ dataset and its subsets. Additionally, when compared to the baseline method, UCF, the CADNet improves the average AUC by 1.11%. Meanwhile, the feature extraction processes of the proposed CADNet on the FF++ dataset are shown in Fig. 4. It shows that the proposed CADNet has the ability to capture the common features of fake images.

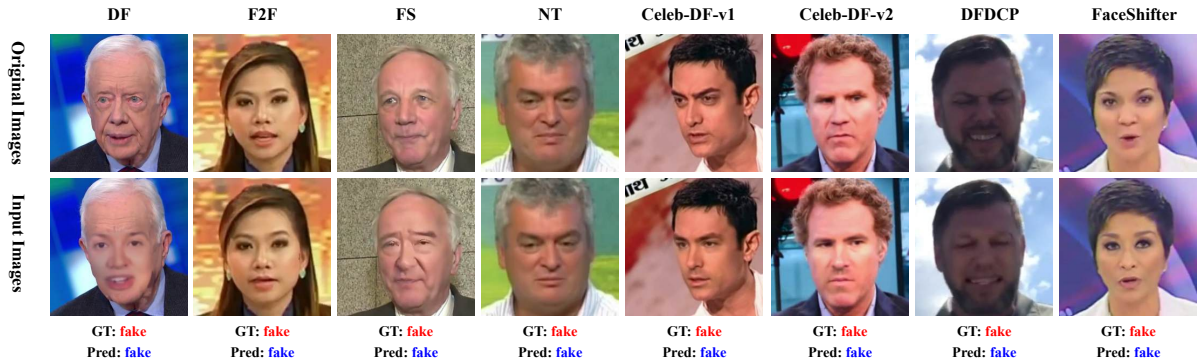

Fig. 5. Visualization of qualitative results on five datasets. The ground truth (GT) and the prediction (Pred) are highlighted in red and blue, respectively.

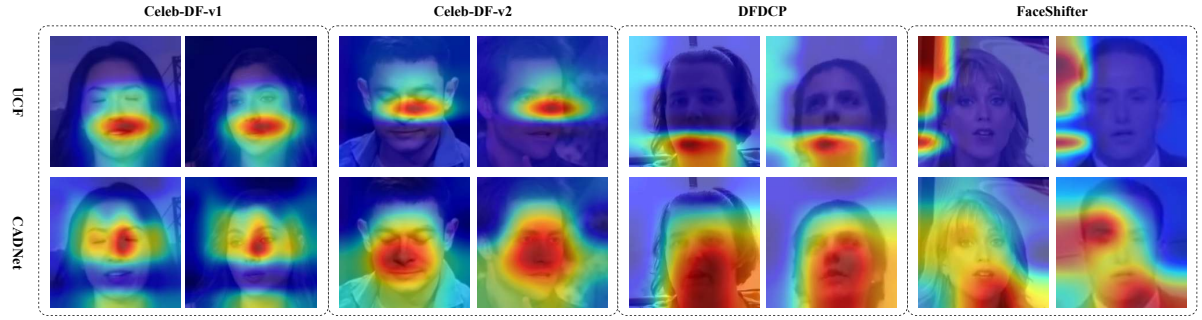

Fig. 6. Attention visualizations of the proposed CADNet and the baseline (UCF). Compared with the baseline, the proposed CADNet focuses on more facial regions.

2) Cross-dataset validation: To evaluate the cross-domain generalization capability of the proposed model, the cross-dataset validation was performed in four unseen out-of-distribution datasets. The framework was trained exclusively on the FF++ dataset and subsequently evaluated across four distinct deepfake datasets: Celeb-DF-v1, Celeb-DF-v2, DFDCP, and FaceShifter. The evaluation results are presented in Table II. Compared to other methods, the proposed model performs best on Celeb-DF-v2, DFDCP, and FaceShifter. It achieves the second-best result on Celeb-DF-v1, only behind SPSL [6].

3) Validation with baseline: To demonstrate the superiority of the proposed method, we compared the proposed CADNet with the baseline method, UCF [5]. Comparative results are presented in Table III. The CADNet outperforms the baseline models across all metrics on the FF++, DF, F2F, and FS datasets. On the NT dataset, CADNet surpasses the baseline in terms of AP and EER.

The subjective results of the CADNet are shown in Fig. 5. It demonstrates the ability of the proposed method to determine the authenticity of images across multiple datasets. Meanwhile, the visualization of the attention map is presented in Fig. 6. Compared to the UCF, the proposed network can focus on a larger range of facial regions in various datasets. Thus, the proposed CADNet is more capable of detecting forgery traces than the UCF model.

TABLE II
COMPARATIVE RESULTS ON CROSS-DATASET VALIDATION. THESE MODELS ARE TRAINED ON THE FF++ DATASET AND TESTED ON THE CELEB-DF-V1, CELEB-DF-V2, DFDCP, AND FACESHIFTER, RESPECTIVELY. THE EVALUATION METRIC IS AUC AND AVERAGE AUC (AVG.). THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Methods	Celeb-DF-v1	Celeb-DF-v2	DFDCP	FaceShifter	Avg.
Meso4 [36]	0.7358	0.6091	0.5994	0.5660	0.6276
MesoIncep [36]	0.7366	0.6966	0.7561	0.6438	0.7083
CNN-Aug [37]	0.7420	0.7027	0.6170	0.5985	0.6651
Xception [31]	0.7794	0.7365	0.7374	0.6249	0.7196
EfficientB4 [38]	0.7909	0.7487	0.7283	0.6162	0.7210
Capsule [39]	0.7909	0.7472	0.6568	0.6465	0.7104
FWA [40]	0.7897	0.6680	0.6375	0.5551	0.6626
X-ray [41]	0.7093	0.6786	0.6942	0.6553	0.6844
FFD [22]	0.7840	0.7435	0.7426	0.6056	0.7189
CORE [21]	0.7798	0.7428	0.7341	0.6032	0.7150
Recce [20]	0.7677	0.7319	0.7419	0.6095	0.7128
UCF [5]	0.7696	0.7391	0.7594	0.6462	0.7286
F3Net [7]	0.7769	0.7352	0.7354	0.5914	0.7097
SPSL [6]	**0.8150**	0.7650	0.7408	0.6437	0.7411
SRM [8]	0.7926	0.7552	0.7408	0.6014	0.7225
CADNet (Ours)	0.7986	**0.7655**	**0.7683**	**0.6882**	**0.7552**

E. Analysis of robustness

Considering the ubiquity of image processing, we investigate the performance under several perturbations, including image compression, Gaussian blur, CLAHE contrast, saturation, and pixelation. The training and testing dataset is the FF++ dataset. In the details of robustness analysis, we introduce the five standardized perturbation methods. **Image compression:** The input images were subjected to lossy compression at a quality factor of 70%. **Gaussian blur:** A

TABLE III
COMPARATIVE RESULTS OF THE PROPOSED CADNET AND THE BASELINE (UCF). THE TRAINING DATASET IS THE FF++ DATASET, AND THE METRICS ARE ACC, AP, AUC, AND EER. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

	FF++				DF				F2F				FS				NT			
	ACC↑	AP↑	AUC↑	EER↓	ACC↑	AP↑	AUC↑	EER↓	ACC↑	AP↑	AUC↑	EER↓	ACC↑	AP↑	AUC↑	EER↓	ACC↑	AP↑	AUC↑	EER↓
Baseline	0.9445	0.9960	0.9705	0.0589	0.9555	0.9943	0.9883	0.0355	0.9445	0.9924	0.9840	0.0388	0.9049	0.9698	0.9896	0.0922	**0.9049**	0.9698	0.9441	0.0922
Ours	**0.9479**	**0.9965**	**0.9854**	**0.0576**	**0.9588**	**0.9951**	**0.9940**	**0.0330**	**0.9600**	**0.9948**	**0.9929**	**0.0299**	**0.9702**	**0.9948**	**0.9933**	**0.0909**	0.8798	**0.9730**	**0.9663**	**0.0909**

TABLE IV
ROBUSTNESS EVALUATION ON FF++ DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Methods	Compress	Blur	Contrast	Saturate	Pixelate
Xception [31]	0.5956	0.6176	0.6110	0.5992	0.5996
CORE [21]	0.5508	0.5136	0.5689	0.5365	0.5372
Recce [20]	0.5280	0.5033	0.5472	0.5354	0.5361
UCF [5]	0.5925	0.6247	0.6187	0.5931	0.6055
SPSL [6]	0.5969	0.6280	0.6334	0.6288	0.6284
SRM [8]	0.5583	0.5487	0.5840	0.5672	0.5691
CADNet (Ours)	**0.6335**	**0.6436**	**0.6673**	**0.6533**	**0.6552**

15×15 Gaussian kernel with standard deviation (σ=1.1) was applied for image blurring. **CLAHE contrast:** Adaptive histogram equalization was performed using an 8×8 grid pattern with a clip limit threshold set to 2.0 for local contrast restriction. **Saturation:** In the HSV color space, the saturation channel was amplified by a factor of 1.1. **Pixelation:** Images were first downsampled to 256×256 resolution and subsequently upsampled to their original dimensions using interpolation. Comparative results are shown in Table IV. The proposed CADNet achieves the best results in image compression, Gaussian blur, CLAHE contrast, saturation, and pixelation operations. It shows that CADNet outperforms other SOTA models when it is subjected to the aforementioned disturbances.

F. Ablation study

To assess the effectiveness of the core components, ablation experiments were conducted on the FF++ dataset. The results are shown in Table V. In ablation studies, the Baseline is the UCF. It proves that when the CSC and FDD modules are equipped individually, the performance in AUC, ACC, AP, and EER can be improved steadily. When the two modules are equipped simultaneously, AUC, ACC, and AP metrics are further enhanced. The value of EER is slightly higher than the "Baseline+CSC" but it is better than the Baseline.

TABLE V
THE ABLATION STUDIES OF THE PROPOSED CADNET. THE AUC, ACC, AP, AND EER ARE USED AS EVALUATION METRICS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Methods	AUC↑	ACC↑	AP↑	EER↓
Baseline	0.9831	0.9445	0.9960	0.0589
Baseline + CSC	0.9842	0.9455	0.9961	**0.0547**
Baseline + FDD	0.9841	0.9460	0.9962	0.0587
Baseline + CSC + FDD	**0.9854**	**0.9479**	**0.9965**	0.0576

V. CONCLUSION

In this work, the CADNet is proposed to enhance the generalization ability of deepfake detection in the facial payment system. In the proposed CADNet, a CSC module is designed to focus on local features while emphasizing non-local facial regions. It can enhance the ability of the network to detect general forgery traces. Meanwhile, an FDD module is designed to decouple high-frequency and low-frequency features for the exploration of frequency domain features. It enables the network to focus on features that are overlooked in the spatial domain. Experimental results demonstrate that the proposed CADNet maintains strong generalization capability. It enhances resistance to deepfake technology in financial transactions while mitigating security threats posed by such attacks. While CADNet excels at image-level detection, its current single-frame analysis paradigm cannot inherently identify temporal artifacts in video forgeries. Future work will extend this framework to video deepfake detection by analyzing inconsistencies between frames.

ACKNOWLEDGEMENTS

This work has been funded by the Shandong Province Undergraduate Teaching Reform Project (No.Z2024184).

CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest related to this publication.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article.

REFERENCES

[1] Q. Li, M. Gao, G. Zhang, W. Zhai, J. Chen, and G. Jeon, "Towards multimodal disinformation detection by vision-language knowledge interaction," *Information Fusion*, vol. 102, p. 102037, February 2024.

[2] G. Zhang, M. Gao, Q. Li, W. Zhai, and G. Jeon, "Multi-modal generative deepfake detection via visual-language pretraining with gate fusion for cognitive computation," *Cognitive Computation*, vol. 16, no. 6, pp. 2953–2966, 2024.

[3] G. Zhang, M. Gao, Q. Li, S. Guo, and G. Jeon, "Detecting sequential deepfake manipulation via spectral transformer with pyramid attention in consumer iot," *IEEE Transactions on Consumer Electronics*, 2024.

[4] M. S. Rana and A. H. Sung, "Deepfakestack: A deep ensemble-based learning technique for deepfake detection," in *2020 7th IEEE international conference on cyber security and cloud computing (CSCloud)/2020 6th IEEE international conference on edge computing and scalable cloud (EdgeCom)*. IEEE, 2020, pp. 70–75.

[5] Z. Yan, Y. Zhang, Y. Fan, and B. Wu, "Ucf: Uncovering common features for generalizable deepfake detection," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 22 355–22 366.

[6] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 772–781.

[7] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Computer Vision – ECCV 2020*. Cham: Springer International Publishing, 2020, pp. 86–103.

[8] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 16 317–16 326.

[9] H. Cheng, Y. Guo, T. Wang, L. Nie, and M. Kankanhalli, "Towards generalizable deepfake detection by primary region regularization," *arXiv preprint arXiv:2307.12534*, 2023.

[10] G. Zhang, M. Gao, Q. Li, W. Zhai, G. Zou, and G. Jeon, "Disrupting deepfakes via union-saliency adversarial attack," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 2018–2026, 2023.

[11] G. Zhang, Q. Li, M. Gao, and G. Jeon, "Towards sequential deepfake detection using deep learning for privacy protection," *IEEE Consumer Electronics Magazine*, vol. 14, pp. 42–48, 2024.

[12] S. Guo, Q. Li, M. Gao, G. Zhang, J. Pan, and G. Jeon, "Deep learning-based face forgery detection for facial payment systems," *IEEE Consumer Electronics Magazine*, vol. 14, no. 3, pp. 80–86, 2024.

[13] S. Guo, Q. Li, M. Gao, G. Zhang, and G. Jeon, "Smart city security: Fake news detection in consumer electronics," *IEEE Consumer Electronics Magazine*, pp. 1–7, 2024.

[14] Z. Yang, J. Liang, Y. Xu, X.-Y. Zhang, and R. He, "Masked relation learning for deepfake detection," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1696–1708, 2023.

[15] M.-Y. Tsai, H.-H. Cho, C.-M. Yu, Y.-C. Chang, and H.-C. Chao, "Effective adversarial examples identification of credit card transactions," *IEEE Intelligent Systems*, vol. 39, no. 4, pp. 50–59, 2024.

[16] M. T. Usman, H. Khan, S. K. Singh, M. Y. Lee, and J. Koo, "Efficient deepfake detection via layer-frozen assisted dual attention network for consumer imaging devices," *IEEE Transactions on Consumer Electronics*, pp. 1–1, 2024.

[17] J. Tian, C. Yu, X. Wang, P. Chen, Z. Xiao, J. Dai, J. Han, and Y. Chai, "Real appearance modeling for more general deepfake detection," in *European Conference on Computer Vision*. Springer Nature Switzerland, 2025, pp. 402–419.

[18] M. Wu, J. Ma, R. Wang, S. Zhang, Z. Liang, B. Li, C. Lin, L. Fang, and L. Wang, "Traceevader: Making deepfakes more untraceable via evading the forgery model attribution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 18, 2024, pp. 19 965–19 973.

[19] S. Guo, Q. Li, M. Gao, X. Zhu, and I. Rida, "Generalizable deepfake detection via spatial kernel selection and halo attention network," *Image and Vision Computing*, vol. 160, p. 105582, July 2025.

[20] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 4113–4122.

[21] Y. Ni, D. Meng, C. Yu, C. Quan, D. Ren, and Y. Zhao, "Core: Consistent representation learning for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2022, pp. 12–21.

[22] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, 2020, pp. 5781–5790.

[23] W. Lu, L. Liu, B. Zhang, J. Luo, X. Zhao, Y. Zhou, and J. Huang, "Detection of deepfake videos using long-distance attention," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[24] Y. Wang, K. Yu, C. Chen, X. Hu, and S. Peng, "Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7278–7287.

[25] S. Das, M. S. Islam, and M. R. Amin, "Gca-net: utilizing gated context attention for improving image forgery localization and detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 81–90.

[26] F. Wang, Q. Chen, B. Jing, Y. Tang, Z. Song, and B. Wang, "Deepfake detection based on the adaptive fusion of spatial-frequency features," *International Journal of Intelligent Systems*, vol. 2024, no. 1, p. 7578036, 2024.

[27] J. Gao, Z. Xia, G. L. Marcialis, C. Dang, J. Dai, and X. Feng, "Deepfake detection based on high-frequency enhancement network for highly compressed content," *Expert Systems with Applications*, vol. 249, p. 123732, September 2024.

[28] J. Liu, J. Wang, P. Zhang, C. Wang, D. Xie, and S. Pu, "Multi-scale wavelet transformer for face forgery detection," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 1858–1874.

[29] M. Wolter, F. Blanke, R. Heese, and J. Garcke, "Wavelet-packets for deepfake image analysis and detection," *Machine Learning*, vol. 111, no. 11, pp. 4295–4327, 2022.

[30] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510.

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

[32] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3207–3216.

[33] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[34] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5074–5083.

[35] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu, "Deepfakebench: A comprehensive benchmark of deepfake detection," in *Advances in Neural Information Processing Systems*, vol. 36. Curran Associates, Inc., 2023, pp. 4534–4565.

[36] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1–7.

[37] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "Cnn-generated images are surprisingly easy to spot... for now," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8695–8704.

[38] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114.

[39] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2307–2311.

[40] Y. Li, "Exposing deepfake videos by detecting face warping artif acts," *arXiv preprint arXiv:1811.00656*, 2018.

[41] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5001–5010.

[42] D. Kinga, J. B. Adam *et al.*, "A method for stochastic optimization," in *International conference on learning representations (ICLR)*, vol. 5. San Diego, California;, 2015, p. 6.

[43] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8024–8035, 2019.