Image and Vision Computing 160 (2025) 105582

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Generalizable deepfake detection via Spatial Kernel Selection and Halo Attention Network

Siyou Guo [a,1], Qilei Li [b,1], Mingliang Gao [a,*], Xianxun Zhu [c], Imad Rida [d]

[a] School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, 255000, China
[b] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, United Kingdom
[c] School of Communication and Information Engineering, Shanghai University, Shanghai, China
[d] Laboratoire Biomécanique et Bioingénierie UMR 7338, Centre de Recherches de Royallieu, Université de Technologie de Compiègne, Compiègne, France

ARTICLE INFO

Keywords:
Privacy protection
Deepfake
Deepfake detection
Deep learning
Attention mechanism

ABSTRACT

The rapid advancement of AI-Generated Content (AIGC) has enabled the unprecedented synthesis of photorealistic facial images. While these technologies offer transformative potential for creative industries, they also introduce significant risks due to the malicious manipulation of visual media. Current deepfake detection methods struggle with unseen forgeries due to their inability to consider the effects of spatial receptive fields and local representation learning. To bridge these gaps, this paper proposes a Spatial Kernel Selection and Halo Attention Network (SKSHA-Net) for deepfake detection. The proposed model incorporates two key modules, namely Spatial Kernel Selection (SKS) and Halo Attention (HA). The SKS module dynamically adjusts the spatial receptive field to capture subtle artifacts indicative of forgery. The HA module focuses on the intricate relationships between neighboring pixels for local representation learning. Comparative experiments on three public datasets demonstrate that SKSHA-Net outperforms the state-of-the-art (SOTA) methods in both intra-testing and cross-testing.

1. Introduction

Deepfake technology advanced rapidly alongside AIGC technologies, such as Generative Adversarial Networks (GANs) [1] and diffusion models [2]. However, deepfake technology has been misused by malicious actors, especially in facial manipulation systems. For instance, deepfake technology is exploited to manipulate facial features in images, altering identities or modifying facial characteristics. The misuse risks a surge in disinformation, which could lead to social unrest and reputational damage [3,4]. As AIGC evolves, it becomes progressively challenging to distinguish between genuine and manipulated counterparts. Consequently, it is crucial to develop effective methods for detecting deepfakes to maintain the integrity of visual media.

Contemporary deepfake detection methods achieve this goal by conducting a binary classification problem, *i.e.*, real or fake, which relies on learning the pattern of deep features extracted by the neural network. Numerous efforts have been proposed to extract distinctive features for deepfake detection from different perspectives. Early deepfake detection methods [5–9] are heavily reliant on non-parametric handcrafted features. These methods focus on aspects of facial features, lighting, and skin texture to analyze possible inconsistencies in deep forgeries directly. Recent studies [10–15] have focused on three well-known approaches for deepfake detection, *i.e.*, anomaly-based methods, pixel and statistical feature-based methods, and generic neural network-based methods. The anomaly-based methods [10,11] are trained only on real images and treat fake images as anomalies in the evaluation process. The pixel and statistical feature-based methods [12–14] extract detailed information from images and videos, *e.g.*, color distribution, textures, and edges. The generic neural network-based methods [16,17] employed pre-trained models to extract high-level features from images or videos for deepfake detection.

Nevertheless, these methods can only conduct binary classification to detect if any manipulation has happened within the image, rather than provide fine-grained manipulation types, as illustrated in Fig. 1. In this regard, numerous recent methods [18–23] have been proposed to improve the generalization ability of the model and the understanding of the unknown forgery patterns. For example, Yin et al. [19] introduced the Domain Adaptive Batch Normalization (DABN) strategy to alleviate the domain distribution gap across diverse datasets.

* Corresponding author.
E-mail addresses: 23504030565@stumail.sdut.edu.cn (S. Guo), q.li@qmul.ac.uk (Q. Li), mlgao@sdut.edu.cn (M. Gao), zhuxianxun@shu.edu.cn (X. Zhu), imad.rida@utc.fr (I. Rida).
[1] Authors contributed equally.

https://doi.org/10.1016/j.imavis.2025.105582
Received 24 February 2025; Received in revised form 14 April 2025; Accepted 7 May 2025
Available online 27 May 2025

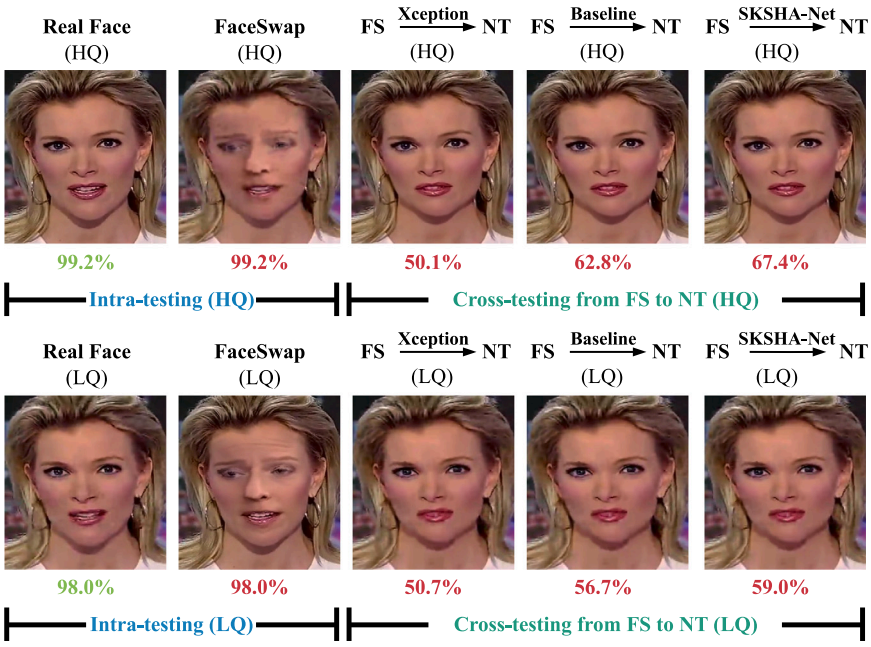

Fig. 1. Detection performance under different protocols. FaceSwap (FS) and NeuralTextures (NT) refer to the types of manipulations. Here, we demonstrate two cases: (1) high-quality deepfakes (top). (2) low-quality (bottom) deepfakes. The proposed SKSHA-Net performs significantly better than the Xception model and baseline on low-quality deepfakes. (Note: Column 2 contains FS-generated data, and columns 3–5 contain NT-generated forgeries.).

Cao et al. [20] proposed the Reconstruction-classification Learning (RECCE) framework by learning differences in real face images through reconstruction. Dong et al. [22] identified that binary classifiers rely on identities present in the training data and designed an ID-unaware deepfake detection model to minimize this dependency. Despite these advances, current deepfake detection methods still face challenges in generalizing unseen manipulation types. While methods like DABN [19] and RECCE [20] improve cross-domain adaptation, they often rely on dataset-specific artifacts or limited spatial representations. As a result, their performance degrades when applied to new manipulation techniques or real-world scenarios where forgery methods evolve rapidly. We identify two major factors limiting generalization: **(1) Fixed receptive fields:** Most detectors use predefined convolutional kernels, which may miss subtle artifacts at varying scales. **(2) Local-context trade-off:** Existing approaches either focus too narrowly on local patches (lacking global consistency) or oversimplify spatial relationships (missing fine-grained traces).

To address the limitations of existing deepfake detection methods in generalizing unseen forgeries, this work proposes a Spatial Kernel Selection and Halo Attention Network (SKSHA-Net). To be specific, the SKS enables the network to extract features at different spatial receptive fields, while the HA mechanism enables the network to focus on the contextual information surrounding each feature, leading to more robust and accurate deepfake detection. To summarize, the main contribution of this paper is as follows:

- We identify the limitation of existing deepfake detection methods in their fixed spatial receptive fields and lack of local contextual learning, which restricts their generalization capabilities. To address this problem, we propose SKSHA-Net for deepfake detection.
- To address the challenge of capturing subtle forgery traces, a Spatial Kernel Selection (SKS) module was built to dynamically adjust the spatial receptive field. To achieve local feature learning interactions, the Halo Attention (HA) mechanism was adopted to focus on contextual relationships of neighboring pixels.
- Comparative experiments on three public datasets demonstrate that SKSHA-Net outperforms the SOTA methods in accuracy and generalization. It can significantly improve the detection of unknown forgeries.

The paper is organized as follows: Section 2 reviews the related work. Section 3 details the proposed model. Section 4 analyses the experimental results. Section 5 shows the ablation results and comprehensive analysis. Section 6 concludes the paper.

2. Related work

2.1. Deepfake detection

Deepfake detection is recognizing and verifying the authenticity of deepfake images or videos. Early deepfake detection methods [24–27] are mainly based on facial cues, *e.g.*, head movements and facial expressions. Li et al. [24] found that deepfake techniques tend to produce subtle facial deformations when synthesizing human faces, and these deformations are characteristic of deepfake. By analyzing these features, real faces and deepfakes can be effectively distinguished. Recently, advanced neural networks have been the predominant approach for detecting deepfakes. For example, Rössler et al. [28] improved deepfake detection performance by retraining an XceptionNet on manipulated face datasets. Zhou et al. [29] proposed a two-stream neural network for deepfake detection. One branch examines visual appearance, while the other focuses on local noise patterns. To improve detection accuracy, Nguyen et al. [30] proposed a capsule network for deepfake detection.

Although these aforementioned methods can achieve high accuracy on the in-domain (same type of manipulation) dataset, the detection performance drops drastically when are tested on an out-of-domain (different type of manipulation) forged dataset [31]. To cope with such uncertain forgery patterns and improve the generalizability, Liu et al. [32] designed residual federated learning and a variational autoencoder to achieve robust and generalizable forgery detection. The RECCE framework [20] was proposed to learn the reconstruction differences between real and fake faces. Nevertheless, the RECCE model falls short in extracting fine-grained features. Therefore, the performance of the RECCE is still unsatisfactory in terms of generalization capabilities.

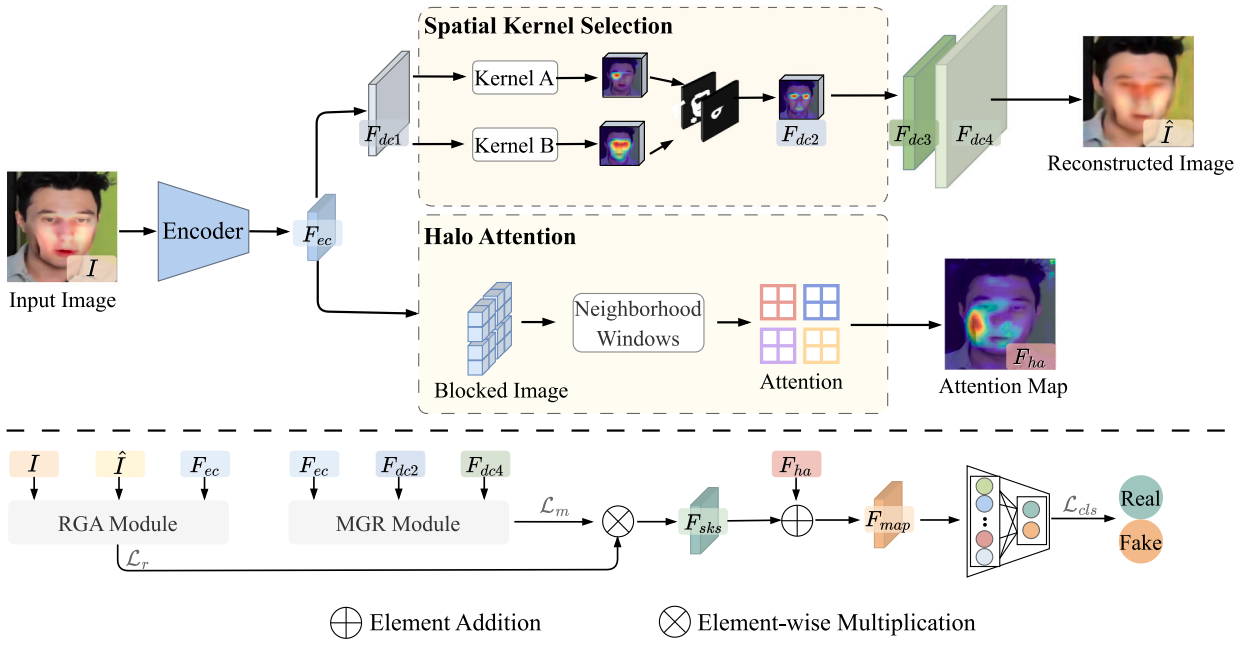

Fig. 2. The framework of SKSHA-Net for deepfake detection. It consists of two parallel pathways: (1) a reconstruction network that processes input image I to generate reconstructed output $\hat{I}$, and (2) a Halo Attention branch that focuses on the contextual relationships of neighboring pixels. .

2.2. Selective kernel mechanism

Deepfake detection methods greatly benefit from effective context modeling due to the subtle and localized nature of manipulations [33]. While traditional generic neural networks with fixed kernel sizes struggle to capture diverse forgery patterns, adaptive kernel size selection provides an alternative and powerful approach to dynamic context modeling [34] with various scales of receptive fields. Recently, several representative selective kernel mechanisms have been proposed. For example, Yang et al. [35] proposed the CondConv that dynamically computes a convolution kernel for each sample, improving the network's performance. Li et al. [36] enabled each neuron to dynamically adjust its receptive field size based on multiple scales of input information by designing a building block called Selective Kernel (SK) unit. Inspired by SK, Self-Calibrated Convolutions (SC) [37] leverages branch attention to capture information and employs spatial attention to refine localization performance. Zhang et al. [38] introduced the ResNeSt, which is a strategy of partitioning the input feature map into multiple groups, thus enhancing the model's ability to capture and represent complex spatial relationships. Li et al. [39] dynamically adjusted the spatial receptive fields of the convolutional kernel to extract features from targets at different scales.

2.3. Attention mechanism

Attention mechanism is an effective tool in the computer vision domain [40,41]. The attention mechanism is inspired by the fact that humans are naturally good at recognizing salient areas in complex scenes. In general, the attention mechanism enables networks to adjust the weight of input features dynamically [42]. Many attention models have been proposed to enhance feature representation. Channel attention mechanisms, such as Squeeze-and-Excitation (SE) [43] and Efficient Channel Attention (ECA) [44], emphasize informative feature channels. Spatial attention mechanisms, like Halo Attention (HA) [45], highlight important spatial regions within the feature maps. Previous works have demonstrated the effectiveness of attention in forgery detection. For example, Su et al. [46] employed the attention mechanism to highlight specific facial regions in each video frame. To solve a fine-grained classification problem in deepfake detection, Zhao et al. [18] employed multiple attention mechanisms to focus on image regions and texture details. Lu et al. [47] employed a long-distance attention mechanism to acquire artifacts from the time and spatial domains. This work introduces the HA to enhance accuracy and investigate the generalization capability of the deepfake detection network.

3. Proposed method

3.1. Overview

The framework of the proposed SKSHA-Net is depicted in Fig. 2. The input image $I \in \mathbb{R}^{C\times H\times W}$ is processed to extract hierarchical features $F_{ec} \in \mathbb{R}^{728\times 19\times 19}$ by an encoder, which is initialized using a pre-trained Xception model. Subsequently, the extracted features are processed in two distinct routes. In the first route, the 128-channel feature maps $F_{enc} \in \mathbb{R}^{128\times 76\times 76}$ are fed into the Spatial Kernel Selection (SKS) module to capture subtle artifacts indicative of forgeries while maintaining the original feature dimensions. This path forms the reconstruction network that generates the reconstructed image $\hat{I}$. The noise introduction technique strategically enlarges the image's coding region, effectively masking out the distorted blank coding points [48]. The reconstruction network equation is formulated as follows:

$$\hat{I} = f_{rec}(\tilde{I}), \tag{1}$$

where the variable $\tilde{I}$ denotes the output obtained by introducing white noise during the training period. The symbol $f_{rec}(\cdot)$ denotes the reconstruction network's processing function.

In the second route, the 728-channel encoder features $F_{ec} \in \mathbb{R}^{728\times 19\times 19}$ are processed by the Halo Attention (HA) module that preserves the input dimensionality. The HA module focuses on the intricate relationships between neighboring pixels for local representation learning. The HA mechanism generates an attention map F_{ha}, which focuses on specific regions within the image that are relevant for detecting forgeries.

Following these steps, SKSHA-Net combines two modules, namely the Reconstruction Guided Attention (RGA) module and the Multiscale Graph Reasoning (MGR) module. The RGA and MGR modules refine and enhance the feature map to obtain F_{sks}. Finally, the feature maps obtained from both routes are aggregated by summing up and serve as

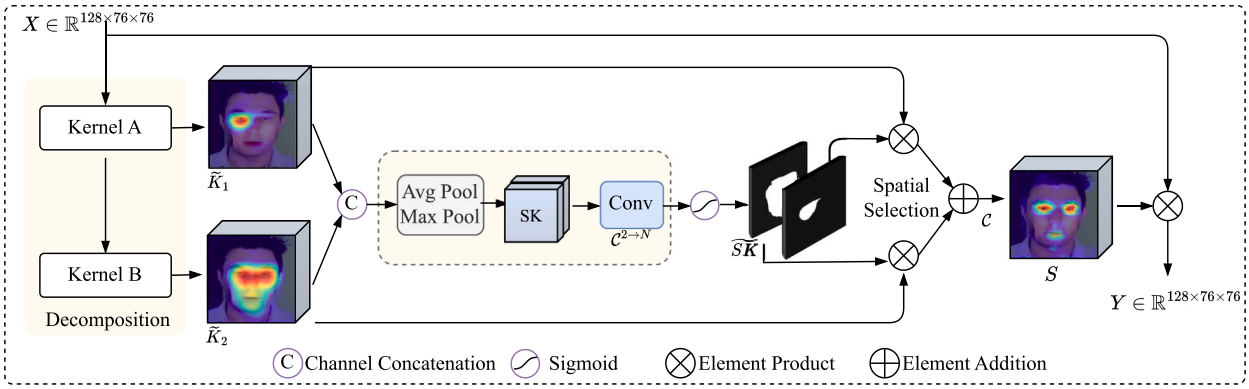

Fig. 3. Schematic diagram of the SKS module.

the foundation for the classifier to discriminate between real and fake images.

3.2. Spatial kernel selection module

To dynamically adjust spatial receptive fields, we employ an SKS mechanism [39] that selectively extracts feature maps from convolutional kernels at various scales. The architecture of the SKS mechanism is shown in Fig. 3.

To enhance the flexibility of the SKS mechanism, a larger kernel convolution is constructed by explicitly decomposing it into a series of depth-wise convolutions. These convolutions grow with increasing kernel size and dilation rates, efficiently capturing long-range contexts. For the ith depth-wise convolution, the receptive field RF_i, dilation rate r_i, and kernel size k_i, are expanded according to the following rules:

$$k_{i-1} \leq k_i; r_1 = 1, r_{i-1} < r_i \leq RF_{i-1}, \tag{2}$$

$$RF_1 = k_1, RF_i = r_i(k_i - 1) + RF_{i-1}. \tag{3}$$

A sequence of depth-wise convolutions with varying receptive fields is employed to extract multi-range contextual features from the input X:

$$K_0 = X, \quad K_{i+1} = F_i^{dw}(U_i), \tag{4}$$

where $F_i^{dw}(\cdot)$ represents depth-wise convolutions. To enable interaction across channels for every spatial feature vector, we decompose the kernels into N components. Each kernel undergoes processing via a 1×1 convolutional layer:

$$\widetilde{K}_i = F_i^{1\times1}(K_i), \quad \text{for} \quad i \in [1, N]. \tag{5}$$

The process commences by extracting features from convolutional kernels with diverse receptive fields. These kernels capture discriminative information at different scales of receptive fields within the input image. We denote these extracted features as $\widetilde{K}_1, \widetilde{K}_2, \ldots, \widetilde{K}_i$. Subsequently, these features are concatenated along the channel dimension to result in a combined feature map denoted as $\widetilde{K}$:

$$\widetilde{K} = [\widetilde{K}_1; \ldots; \widetilde{K}_i]. \tag{6}$$

In the SKS module, to efficiently extract the salient spatial relationships within the combined feature map $\widetilde{K}$, we incorporate two candidate operators:

(1) Channel-based Average Pooling ($\mathcal{P}_{avg}(\cdot)$): This operation calculates the average value for each spatial location across all channels in $\widetilde{K}$.

(2) Channel-based Maximum Pooling ($\mathcal{P}_{max}(\cdot)$): This operation identifies the maximum value for each spatial location across all channels in $\widetilde{K}$.

The results of these pooling operations are represented by two groups of feature maps, namely SK_{avg} and SK_{max}:

$$SK_{\text{avg}} = \mathcal{P}_{avg}(\widetilde{K}), SK_{\text{max}} = \mathcal{P}_{max}(\widetilde{K}). \tag{7}$$

These features are then aggregated by concatenating along the channel dimension to enable interaction between the information captured through average and maximum pooling, which is further refined by a convolution layer denoted as $\mathcal{C}^{2\to N}(\cdot)$. This layer has two input channels (corresponding to the average and maximum pooled features) and outputs N spatial attention maps.

$$\widehat{SK} = \mathcal{C}^{2\to N}([SK_{avg}; SK_{max}]). \tag{8}$$

This convolution transforms the combined spatial information from average and maximum pooling into N distinct spatial attention maps. These maps will later selectively focus on relevant regions in the input features.

Each of N spatial attention maps produced in the previous step $SK_1, SK_2, \ldots, SK_i$ is passed through a sigmoid activation function $\sigma(\cdot)$:

$$\widetilde{SK}_i = \sigma(\widehat{SK}_i). \tag{9}$$

The sigmoid function squashes the values in each attention map to a range between 0 and 1. As a result, each spatial location within an attention map will have a value close to 0 (indicating low importance) or close to 1 (indicating high importance).

Following spatial selection, the features from each decomposed large kernel, $\widetilde{K}_1, \widetilde{K}_2, \ldots, \widetilde{K}_i$, are element-wise multiplied with their corresponding spatial selection masks $SK_1, SK_2, \ldots, SK_i$. This multiplication scales the features based on their importance in different spatial locations. Regions identified as less relevant by the selection masks contribute less to the final output. Subsequently, the element-wise multiplied feature maps are summed together by another convolution layer $C(\cdot)$. This layer integrates information from all decomposed kernels while incorporating the spatial weighting provided by the selection masks. This process is formulated as:

$$S = C(\sum_{i=1}^{N}(\widetilde{SK}_i \cdot \widetilde{K}_i)). \tag{10}$$

The final output of the SKS denoted as Y, is obtained by performing the element-wise product between the input feature X and the combined attention feature S.

$$Y = S \odot X. \tag{11}$$

The SKS mechanism dynamically modulates the input feature based on the learned spatial weights: Regions with higher selection mask values (scores) have a more dominant influence on the final output, while less relevant areas will be suppressed to avoid interference.

Fig. 4. Schematic diagram of the HA module.

3.3. Halo attention module

Different from conventional self-attention mechanisms, Halo attention (HA) module [45] concentrates on a small localized window around each pixel. This structure enables the proposed framework to capture useful relationships between adjacent pixels. The architecture of the Halo attention is shown in Fig. 4.

Specifically, the input feature map is partitioned into four equally sized blocks. Each block is subjected to a fill operation, *i.e.*, n layers of halos are added around each block. This augmentation expands the perceptual field of each block, resulting in a larger perceptual field for each. The process of Halo attention is formulated as follows:

$$F_{ha} = \text{softmax}(q * k + \text{halo}_{\text{size}}(q, k)) * v, \tag{12}$$

where F_{ha} is the output, q is the query vector, v is the value vector, and k is the key vector. $\text{halo}_{\text{size}}(q, k)$ is the Halo function to compute the Halo size. Subsequently, each block is individually sampled. Ultimately, the feature map is output and combined with the initial input feature map. This seamless integration is facilitated by utilizing the residual jump connection, which effectively amalgamates local and global details.

3.4. Loss function

The proposed model is trained in an end-to-end manner, guided by a combination of three loss functions. These include a reconstruction loss ($\mathcal{L}_{\text{r}}$), a classification loss ($\mathcal{L}_{\text{cls}}$), and a metric learning loss ($\mathcal{L}_{\text{m}}$). The reconstruction loss $\mathcal{L}_{\text{r}}$ is denoted as:

$$\mathcal{L}_r = \frac{1}{N} \sum_{i=1}^{N} \sum_{d=1}^{D} |x_{i,d} - \hat{x}_{i,d}|, \tag{13}$$

where N signifies authentic samples, and D signifies feature dimension. $x_{i,d}$ is the estimated value of the dth feature of the ith authentic sample, and $\hat{x}_{i,d}$ denotes the corresponding feature after reconstruction.

A metric learning loss $\mathcal{L}_{\text{m}}$ minimizes the distance between real data and maximizes the separation between real and synthetic representations in the embedding space. The metric learning loss $\mathcal{L}_{\text{m}}$ is defined as:

$$\mathcal{L}_{\text{m}} = \frac{1}{|R|} \sum_{i,j \in N} d(I_i, I_j) - \frac{1}{|N \times F|} \sum_{i \in N, j \in F} d(I_i, I_j), \tag{14}$$

where N signifies authentic samples and F signifies fake samples. $|N|$ denotes the cardinality of N. $|N \times F|$ denotes the number of elements of the Cartesian product of a set N and a set F. The cosine distance function $d(\cdot, \cdot)$ calculates the pairwise distance between data points. The function is expressed as:

$$d(\alpha, \beta) = \frac{1 - \frac{\alpha}{\|\alpha\|_2} \cdot \frac{\beta}{\|\beta\|_2}}{2}. \tag{15}$$

The loss function $\mathcal{L}_{\text{cls}}$ used for two-class classification is formulated as:

$$\mathcal{L}_{\text{cls}} = -y * log(p) - (1 - y) * log(1 - p), \tag{16}$$

where y denotes the true label of the sample, with 0 denoting the negative class and 1 denoting the positive class. p denotes the probability that the sample predicted by the model belongs to the positive class.

The final loss function of the proposed method is defined as:

$$\mathcal{L} = w_1 \mathcal{L}_{\text{m}} + w_2 \mathcal{L}_{\text{r}} + \mathcal{L}_{\text{cls}}, \tag{17}$$

where the values of w_1 and w_2 determine the importance of each loss term in influencing the training process.

4. Experimental results and analysis

4.1. Experimental setup

Datasets. We conducted extensive experiments on three benchmark datasets, namely FaceForensics++ (FF++) [28], Celeb-DF [49], and WildDeepfake [50]. FF++ is manipulated with four face manipulation approaches: Face2Face (F2F), NeuralTextures (NT), Deepfakes (DF), and FaceSwap (FS). Celeb-DF comprises 590 high-resolution videos meticulously chosen from YouTube celebrity interviews. Compared to FF++, videos in Celeb-DF are more challenging due to the superior quality of the videos and the utilization of more sophisticated deepfake synthesis algorithms. WildDeepfake contains 1,000 videos captured in the real world. The videos in WildDeepfake are more realistic than those in FF++ and Celeb-DF because they are not created in a controlled laboratory environment. Since the raw data is in video format, the face images are extracted from the sequence by RetinaFace [51] technique.

Evaluation Metrics. We employed two metrics to evaluate the proposed SKSHA-Net framework's performance: Accuracy (Acc) and Area Under the Curve (AUC). The Acc is the proportion of correctly predicted samples to the total number of samples. The AUC is a more robust metric unaffected by class imbalance. Higher Acc and AUC indicate better discrimination ability of the model. The Equal Error Rate (EER) was employed to assess the proposed model's performance in cross-dataset experiments. The EER is the threshold at which the false acceptance rate (FAR) and the false rejection rate (FRR) are equal. The network with a lower EER generally performs better.

Implementation Details. We trained the model with 32 samples per batch for 40 iterations. The learning rate is dynamically adjusted during training using a step-learning rate scheduler. A grid search strategy was utilized to optimize key hyperparameters. The Adam optimizer was adopted with a learning rate of 2e-4 and weight decay of 1e-5. The model was trained using PyTorch with two 3090 Ti GPUs in parallel.

Table 1
Objective result of intra-testing on the FF++, WildDeepfake, and Celeb-DF datasets. The best results are highlighted in **bold**.

Methods	FF++ (HQ)		FF++ (LQ)		WildDeepfake		Celeb-DF	
	Acc (%)	AUC (%)	Acc (%)	AUC (%)	Acc (%)	AUC (%)	Acc (%)	AUC (%)
MesoNet [52]	83.10	–	70.47	–	–	–	64.47	–
SPSL [53]	91.50	95.32	81.57	82.82	–	–	–	–
RFM [54]	95.69	98.79	87.06	89.83	77.38	83.92	97.96	99.94
Freq-SCL [55]	96.69	99.28	89.00	92.39	–	–	–	–
Multi-task [56]	85.65	85.43	81.30	75.59	–	–	–	–
Face X-ray [57]	–	87.84	–	61.60	–	–	–	–
Xception [58]	95.73	96.30	86.86	89.30	77.25	86.76	97.90	99.73
Add-Net [50]	96.78	97.74	87.50	91.01	76.25	86.17	96.93	99.55
Two-branch [59]	96.43	98.70	86.34	86.59	–	–	–	–
DFF+LEA [60]	94.14	98.78	79.58	89.93	–	–	–	–
MultiAtt [18]	**97.60**	99.29	88.69	90.40	82.86	90.71	97.92	99.94
RECCE [20]	97.06	**99.32**	91.03	95.02	84.26	91.54	98.59	99.94
SKSHA-Net (Ours)	97.16	**99.32**	**91.32**	**95.21**	**84.72**	**92.62**	**99.29**	**99.95**

Table 2
Objective result of cross-testing. "↓" means the lower the better, and "↑" means the higher the better. The best results are highlighted in **bold**.

Methods	WildDeepfake		Celeb-DF	
	AUC (%)↑	EER (%)↓	AUC (%)↑	EER (%)↓
Xception [58]	62.72	40.65	61.80	41.73
RFM [54]	57.75	45.45	65.63	38.54
Add-Net [50]	62.35	41.42	65.29	38.90
F^3-Net [62]	57.10	45.12	61.51	42.03
MultiAtt [18]	59.74	43.73	67.02	37.90
RECCE [20]	64.31	40.53	68.71	35.73
SKSHA-Net (Ours)	**69.58**	**35.31**	**70.54**	**34.75**

4.2. Experimental results

Intra-testing. Intra-testing thoroughly compares the proposed method against the state-of-the-art (SOTA) approaches. As shown in Table 1, the proposed method performs best on FF++ datasets. On FF++ (HQ) dataset and FF++ (LQ) dataset, the proposed method achieves 97.16 and 99.32, 91.32 and 95.21 in Acc and AUC, respectively. Compared to the RECCE [20], the proposed method achieves comparable performance on the FF++ (HQ) dataset. On the FF++ (LQ) dataset, the proposed SKSHA-Net improves the Acc and AUC metrics by 0.29% and 0.19%, respectively. Given that the MultiAtt [18] is based on the EfficientNet-b4 [61] backbone and achieves the highest accuracy on FF++ (HQ), the Xception-based approach still achieves comparable results. Moreover, the proposed method enhanced detection performance on the realistic WildDeepfake and Celeb-DF datasets.

Cross-testing. Cross-testing was conducted to evaluate the generalization performance of the proposed method to unseen forgeries. The proposed method was trained on the low-quality FF++ dataset and tested on both the WildDeepfake and Celeb-DF datasets, which involve different deepfake generation techniques. The results of the cross-testing experiment are shown in Table 2. On the WildDeepfake dataset, the proposed model outperforms all other competitors in AUC and the EER. Specifically, compared with the second-best method RECCE [20], the proposed SKSHA-Net improves the EER and AUC by 5.22% and 5.27%, respectively. Similarly, on the Celeb-DF dataset, the proposed method ranks first among other methods. It scores 70.54 and 34.75 in AUC and EER, respectively. These experimental results demonstrate the proposed framework's robust generalization ability.

We assessed the generalization ability of the proposed method in detecting deepfakes generated by different operational techniques. Fine-grained cross-testing was performed, *i.e.*, specific manipulation techniques were trained and tested on datasets forged by other manipulation techniques in the FF++ (LQ) dataset. Comparison of generalization results across manipulation methods is shown in Table 3. It shows that the proposed model is generally superior to other methods in terms of unseen forgery techniques. Unlike existing methods that are limited by fixed receptive fields, the SKS module reduces the dependence on specific forgery artifacts by dynamically fusing multi-scale features to enhance generalization. The HA module further improves cross-technique robustness by modeling localized inconsistencies, which often persist across different manipulation methods. This dual-adaptive design enhances the model's generalization when encountering unseen deepfake variations.

5. Discussion

5.1. Ablation study

To evaluate the individual contributions of each component within the SKSHA-Net framework, we conducted a comprehensive ablation study on the WildDeepfake [50] dataset. Comparative results are presented in Table 4. In this analysis, the "Baseline" refers to the reconstruction-classification learning (RECCE) [20]. The "SKS" denotes the Spatial Kernel Selection, and the "HA" represents the Halo Attention module. The ablation study reveals that while incorporating SKS or HA independently leads to modest improvements in Acc and AUC, their combination delivers the most significant performance gains. This highlights the complementary synergy between SKS and HA, which together drive superior detection accuracy. Specifically, SKS captures the global context, while HA focuses on localized details. This balance between broad spatial understanding and pixel-level attention improves detection accuracy.

5.2. Visualization analysis

To enhance the interpretability of the proposed model, we employ Grad-CAM [63] to generate attention maps that highlight the image regions most influential in deepfake detection. These visualizations reveal distinct decision-making patterns across different architectures. The visualization of attention maps under different protocols is shown in Fig. 5.

Baseline: The attention map of the baseline model predominantly focuses on the central facial features, notably the nose and eyes. This indicates a strong reliance on these features for forgery detection.

Baseline+SKS: The attention map of the baseline+SKS model also focuses on the central area of the face but extends to areas beyond those focused on by the baseline. This demonstrates that the baseline+SKS model considers a broader range of features when making forgery detection decisions.

Baseline+HA: Compared to both the baseline model and the baseline+SKS model, the baseline+HA model demonstrates stronger attention to local features in its attention maps.

SKSHA-Net: The attention map of the SKSHA-Net model is the most comprehensive among the four models. This indicates that the SKSHA-Net model considers the most global features.

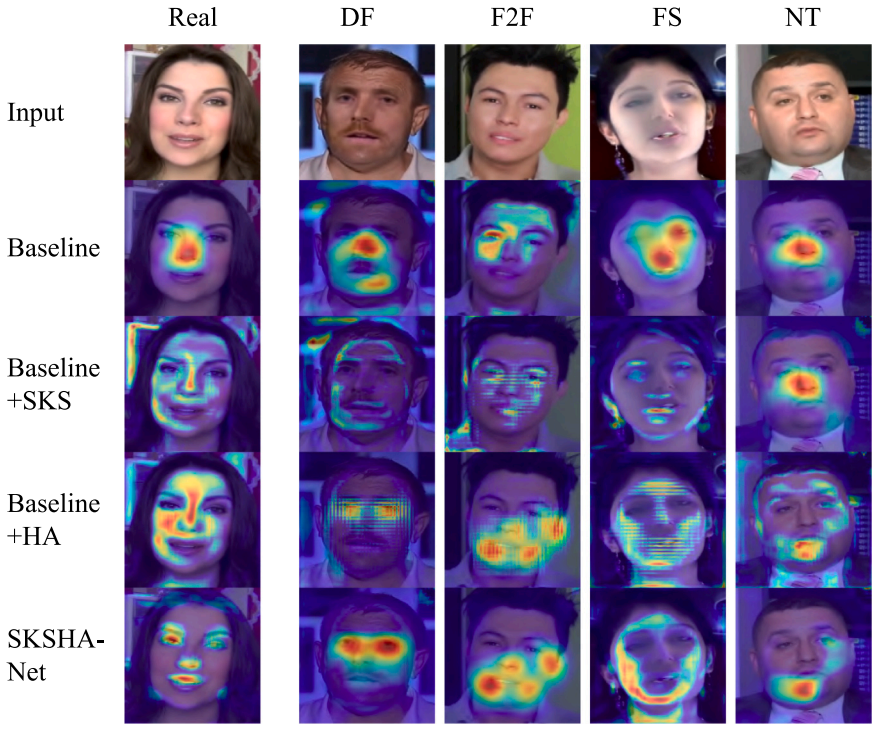

Fig. 5. The visualization of attention maps under different protocols.

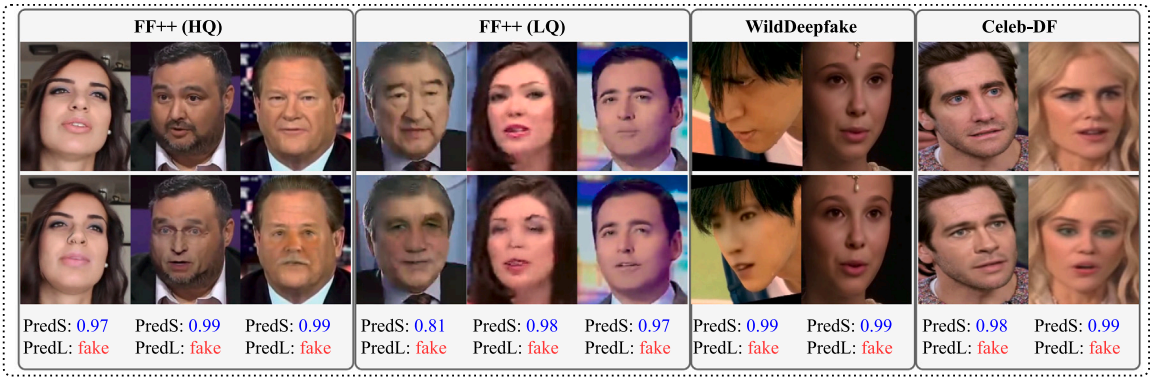

Fig. 6. The inference results of deepfake detection by the proposed model in different datasets. The top row displays unmanipulated face images, and the bottom row shows forged images used as inputs to the inference model. Prediction Score (PredS) indicates the probability of the image being fake. Predicted Label (PredL) denotes the result inferred by the model.

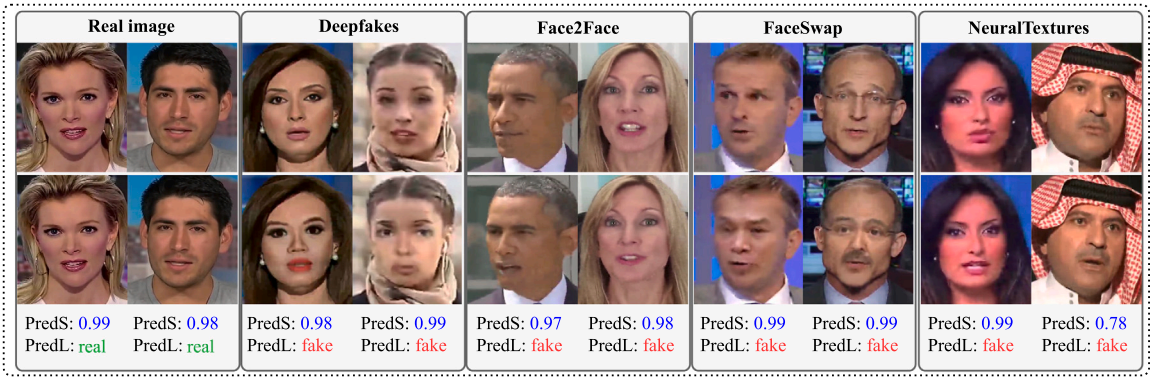

Fig. 7. Inference results of the proposed framework for different manipulation techniques in the FF++ dataset. The top row displays unmanipulated face images, and the bottom row shows forged images used as inputs to the inference model. Prediction Score (PredS) indicates the probability of the image being fake. Predicted Label (PredL) denotes the result inferred by the model.

Table 3
Comparison of generalization results across manipulation methods detection on the FF++ (LQ) dataset. The gray areas indicate Intra-testing. The evaluation metric is AUC (%). The best results are highlighted in **bold**. "Cross Avg" indicates the average value of the Cross-testing.

Methods	Training dataset	Evaluation datasets				
		Deepfakes	Face2Face	FaceSwap	NeuralTextures	Cross Avg.
EfficientNet [61]	Deepfakes	98.63	60.55	59.65	60.87	60.36
GramNet [27]		98.19	57.05	62.80	57.04	58.96
DFF+LEA [60]		98.74	62.17	64.10	61.84	62.70
Freq-SCL [55]		98.91	58.90	66.87	63.61	63.13
MultiAtt [18]		99.51	66.41	67.33	66.01	66.58
RECCE [20]		99.65	70.66	**74.29**	67.34	70.76
SKSHA-Net (Ours)		**99.68**	**72.48**	71.48	**70.31**	**71.42**
EfficientNet [61]	Face2Face	70.60	95.10	58.21	61.74	63.52
GramNet [27]		61.26	92.77	54.19	58.73	58.06
DFF+LEA [60]		67.30	95.41	58.45	59.36	61.70
Freq-SCL [55]		67.55	93.06	55.35	66.66	63.19
MultiAtt [18]		73.04	97.96	65.10	71.88	70.01
RECCE [20]		**75.99**	98.06	64.53	72.32	70.95
SKSHA-Net (Ours)		75.64	**98.09**	**64.92**	**73.38**	**71.31**
EfficientNet [61]	FaceSwap	71.90	59.79	97.98	51.04	60.91
GramNet [27]		73.43	57.56	96.15	51.93	60.97
DFF+LEA [60]		77.73	**56.55**	98.15	53.20	62.49
Freq-SCL [55]		75.90	54.64	98.37	49.72	60.09
MultiAtt [18]		82.33	61.65	98.82	54.79	66.26
RECCE [20]		83.39	**64.44**	98.82	56.70	67.84
SKSHA-Net (Ours)		**83.42**	64.16	**98.89**	**59.03**	**68.87**
EfficientNet [61]	NeuralTextures	79.09	74.21	53.99	88.54	69.10
GramNet [27]		74.56	80.61	60.90	93.34	72.02
DFF+LEA [60]		78.83	80.89	63.70	93.63	74.47
Freq-SCL [55]		79.09	74.21	53.99	88.54	69.10
MultiAtt [18]		74.56	80.61	60.90	93.34	72.02
RECCE [20]		78.83	80.89	63.70	93.63	74.47
SKSHA-Net (Ours)		**79.56**	**81.67**	**64.20**	**93.78**	**75.14**

Table 4
Ablation studies on the critical modules in SKSHA-Net. The best results are highlighted in **bold**.

Methods			Evaluation metrics	
Baseline	SKS	HA	Acc (%)	AUC (%)
✓			84.26	91.54
✓	✓		84.40	91.14
✓		✓	83.44	91.58
✓	✓	✓	**84.72**	**92.62**

From the aforementioned analysis, the limited attention range of the baseline may make it more susceptible to forgeries that are not located in the central region of the face. The Baseline+SKS and Baseline+HA models address this limitation by considering a wider range of features. The SKSHA-Net model takes this even further by considering the most global features. Overall, the SKSHA-Net is the most robust forgery detection model among the four options. This is due to its capability to incorporate the widest range of features, from localized to global.

5.3. Examples and analysis

To validate the detection capabilities of the framework, the inference results of deepfake detection by the proposed framework are illustrated in Figs. 6 and 7. These figures highlight the effectiveness of SKSHA-Net in identifying forgery techniques at a fine-grained level. Fig. 6 showcases the inference results across different datasets. The diversity of datasets underscores the robustness and generalizability of the proposed SKSHA-Net. The model consistently delivers high prediction scores and accurately predicts the authenticity labels of faces. Fig. 7 presents the inference results of various facial manipulation techniques on the FF++ dataset. The proposed SKSHA-Net achieves impressive prediction scores across different manipulation methods, which demonstrates its capability to detect subtle and complex forgeries.

5.4. Analysis of robustness

To evaluate the robustness of the proposed framework, we tested its resilience against common social media image transformations. The experiments simulated five realistic perturbations [64,65], including image compression, contrast jitter, Gaussian blur, saturation jitter, and pixelation. As shown in Table 5, the SKSHA-Net demonstrates superior robustness compared to existing methods when subjected to common image perturbations. Two observations can be made: (i) Existing methods exhibit marked sensitivity to structural degradations, *i.e.*, Gaussian blur (disrupting frequency statistics) and pixelation (eroding texture patterns), with performance drops reaching 4.2%; (ii) Compared to RECCE, SKSHA-Net exhibits consistent robustness. The AUC improves by 2.16% against Gaussian blur and 1.48% against pixelation. The proposed framework improves the AUC by an average of 1.21% across all perturbation tests. These results validate that SKSHA-Net is sufficient to resist common image perturbations with high robustness.

5.5. Failure cases

The Grad-CAM [63] method was employed to analyze instances where the detection model struggled. As illustrated in Fig. 8, the left examples highlight a critical issue, *i.e.*, the model focuses too much on the margin regions of the images and thus draws wrong conclusions. Furthermore, the right examples show that complex backgrounds affect the model's performance. These findings reveal fundamental challenges in real-world deepfake detection, where border artifacts and background clutter frequently interfere with authentic manipulation cues. Future work is expected to mitigate border dependency and background sensitivity via attention constraints and adversarial context augmentation.

5.6. Future directions

Although deep learning forgery detection methods that rely on feature representations have made great strides in recent years, these methods do not consider the role of spatial receptive fields and local representation learning. Therefore, we have developed a generalized forgery detection task to identify forgery clues across different domains, including unknown artifact types. This section addresses several unresolved issues and suggests potential future research directions to inspire the community of researchers.

Table 5
Robustness evaluation in terms of AUC (%) on WildDeepfake dataset. "Avg." represents the average score.

Methods	Image compression	Contrast jitter	Gaussian blur	Saturation jitter	Pixelation	Avg.
Xception [58]	86.01	81.90	78.29	84.96	66.24	79.48
RFM [54]	83.74	79.77	75.34	82.59	71.25	78.54
Add-Net [50]	83.34	84.46	79.66	85.13	64.33	79.38
F^3-Net [62]	86.71	86.53	78.99	87.67	73.23	82.63
MultiAtt [18]	89.64	89.30	80.98	90.37	79.44	85.95
RECCE [20]	89.65	91.19	87.29	91.74	83.88	88.75
SKSHA-Net (Ours)	**90.12**	**92.57**	**89.45**	**92.30**	**85.36**	**89.96**

Fig. 8. Attention maps for failure cases on the WildDeepfake dataset. The three sets of images on the left display the attention the model gives to the margins, while the images on the right highlight the attention given to the complex background.

Robust Deepfake Detection Representation: Deployment of deepfake detection frameworks in the real world faces various challenges, including variations in occlusion, noise, illumination and similar factors. At the same time, malicious adversarial attacks might prevent the detection model from making correct decisions. In the context of security and privacy protection, a crucial objective is to establish a robust detection representation that can effectively resist adversarial samples and noise.

Weakly/Unsupervised Supervised Deepfake Detection: Fully supervised learning is the predominant approach for training current deepfake detection models. However, as deep generative models continue to evolve, more unknown artifact patterns are generated. It is becoming impractical to label every artifact type accurately. A potential solution is to focus on learning generic forgery cues to enhance model performance. Weakly or unsupervised learning techniques (*e.g.*, dynamic convolution, meta-learning, self-attention mechanisms) provide viable approaches to distinguish critical features between authentic and forged images.

Interpretable Deepfake Detection: Interpretable deepfake detection is a crucial area in digital forensics, which focuses on identifying and localizing tampered or forged areas in an image. The challenge is to build highly accurate deepfake detectors with interpretable decision-making. Feature importance analysis and interpretation of cross-modal data may offer viable solutions to address the interpretability problem.

6. Conclusion

The fundamental challenge in deepfake detection comes from learning generalized representations that transcend specific manipulation patterns. Through systematic analysis, we identify two critical limitations in existing approaches, *i.e.*, rigid spatial feature extraction constrained by fixed receptive fields, and insufficient modeling of local artifact interdependencies. To bridge these gaps, this paper proposes a Spatial Kernel Selection and Halo Attention Network (SKSHA-Net) for deepfake detection. A spatial kernel selection module is built to adapt the spatial receptive field to capture forged traces. Meanwhile, Halo attention is adopted to capture useful relationships between adjacent pixels. Experiment results demonstrate the superior performance of the proposed method in deepfake detection. It particularly excels on datasets with unknown forgery patterns.

CRediT authorship contribution statement

Siyou Guo: Writing – original draft, Visualization, Data curation. **Qilei Li:** Validation, Investigation. **Mingliang Gao:** Project administration, Methodology, Investigation. **Xianxun Zhu:** Investigation, Data curation. **Imad Rida:** Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

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

Data availability

The authors do not have permission to share data.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.
[2] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.
[3] M. Chen, X. Liao, M. Wu, PulseEdit: Editing physiological signals in facial videos for privacy protection, IEEE Trans. Inf. Forensics Secur. 17 (2022) 457–471.
[4] S. Guo, Q. Li, M. Gao, G. Zhang, G. Jeon, Smart city security: Fake news detection in consumer electronics, IEEE Consum. Electron. Mag. (2024).
[5] X. Yang, Y. Li, S. Lyu, Exposing deep fakes using inconsistent head poses, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2019, pp. 8261–8265.
[6] F. Matern, C. Riess, M. Stamminger, Exploiting visual artifacts to expose deepfakes and face manipulations, in: 2019 IEEE Winter Applications of Computer Vision Workshops, WACVW, IEEE, 2019, pp. 83–92.
[7] G. Li, Y. Cao, X. Zhao, Exploiting facial symmetry to expose deepfakes, in: 2021 IEEE International Conference on Image Processing, ICIP, IEEE, 2021, pp. 3587–3591.

[8] S. Guo, M. Gao, Q. Li, G. Jeon, D. Camacho, Deepfake detection via a progressive attention network, in: 2024 International Joint Conference on Neural Networks, IJCNN, IEEE, 2024, pp. 1–6.

[9] G. Zhang, M. Gao, Q. Li, S. Guo, G. Jeon, Detecting sequential deepfake manipulation via spectral transformer with pyramid attention in consumer IoT, IEEE Trans. Consum. Electron. (2024).

[10] R. Wang, F. Juefei-Xu, L. Ma, X. Xie, Y. Huang, J. Wang, Y. Liu, FakeSpotter: a simple yet robust baseline for spotting AI-synthesized fake faces, in: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, 2021, pp. 3444–3451.

[11] A. Khodabakhsh, C. Busch, A generalizable deepfake detector based on neural conditional distribution modelling, in: 2020 International Conference of the Biometrics Special Interest Group, BIOSIG, IEEE, 2020, pp. 1–5.

[12] M. Koopman, A.M. Rodriguez, Z. Geradts, Detection of deepfake video manipulation, in: The 20th Irish Machine Vision and Image Processing Conference, IMVIP, 2018, pp. 133–136.

[13] K. Liu, I. Perov, D. Gao, N. Chervoniy, W. Zhou, W. Zhang, Deepfacelab: Integrated, flexible and extensible face-swapping framework, Pattern Recognit. 141 (2023) 109628.

[14] Z. Yu, X. Li, J. Shi, Z. Xia, G. Zhao, Revisiting pixel-wise supervision for face anti-spoofing, IEEE Trans. Biom. Behav. Identity Sci. 3 (3) (2021) 285–295.

[15] B. Liu, B. Liu, M. Ding, T. Zhu, X. Yu, TI2Net: temporal identity inconsistency network for deepfake detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 4691–4700.

[16] Q. Li, M. Gao, G. Zhang, W. Zhai, J. Chen, G. Jeon, Towards multimodal disinformation detection by vision-language knowledge interaction, Inf. Fusion 102 (2024) 102037.

[17] G. Zhang, M. Gao, Q. Li, W. Zhai, G. Jeon, Multi-modal generative DeepFake detection via visual-language pretraining with gate fusion for cognitive computation, Cogn. Comput. (2024) 1–14.

[18] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, N. Yu, Multi-attentional deepfake detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2185–2194.

[19] Z. Yin, J. Wang, Y. Ding, Y. Xiao, J. Guo, R. Tao, H. Qin, Improving generalization of deepfake detection with domain adaptive batch normalization, in: Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia, 2021, pp. 21–27.

[20] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, 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, 2022, pp. 4113–4122.

[21] F. Khalid, A. Javed, H. Ilyas, A. Irtaza, et al., DFGNN: An interpretable and generalized graph neural network for deepfakes detection, Expert Syst. Appl. 222 (2023) 119843.

[22] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, Z. Ge, Implicit identity leakage: The stumbling block to improving deepfake detection generalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3994–4004.

[23] Q. Lv, Y. Li, J. Dong, S. Chen, H. Yu, H. Zhou, S. Zhang, DomainForensics: Exposing face forgery across domains via bi-directional adaptation, IEEE Trans. Inf. Forensics Secur. 19 (2024) 7275–7289.

[24] Y. Li, S. Lyu, Exposing deepfake videos by detecting face warping artifacts, 2018, arXiv preprint arXiv:1811.00656.

[25] A. Malik, M. Kuribayashi, S.M. Abdullahi, A.N. Khan, DeepFake detection for human face images and videos: A survey, Ieee Access 10 (2022) 18757–18775.

[26] Y. Zhao, W. Ge, W. Li, R. Wang, L. Zhao, J. Ming, Capturing the persistence of facial expression features for deepfake video detection, in: Information and Communications Security: 21st International Conference, ICICS 2019, Beijing, China, December 15–17, 2019, Revised Selected Papers 21, Springer, 2020, pp. 630–645.

[27] Z. Liu, X. Qi, P.H. Torr, Global texture enhancement for fake face detection in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8060–8069.

[28] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1–11.

[29] P. Zhou, X. Han, V.I. Morariu, L.S. Davis, Two-stream neural networks for tampered face detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, IEEE, 2017, pp. 1831–1839.

[30] H.H. Nguyen, J. Yamagishi, I. Echizen, Use of a capsule network to detect fake images and videos, 2019, arXiv preprint arXiv:1910.12467.

[31] M. Masood, M. Nawaz, K.M. Malik, A. Javed, A. Irtaza, H. Malik, Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward, Appl. Intell. 53 (4) (2023) 3974–4026.

[32] D. Liu, Z. Dang, C. Peng, Y. Zheng, S. Li, N. Wang, X. Gao, FedForgery: generalized face forgery detection with residual federated learning, IEEE Trans. Inf. Forensics Secur. 18 (2023) 4272–4284.

[33] Y. Huang, F. Juefei-Xu, Q. Guo, Y. Liu, G. Pu, Fakelocator: Robust localization of gan-based face manipulations, IEEE Trans. Inf. Forensics Secur. 17 (2022) 2657–2672.

[34] Y. He, X. Jin, Q. Jiang, Z. Cheng, P. Wang, W. Zhou, LKAT-GAN: A GAN for thermal infrared image colorization based on large kernel and AttentionUNet-transformer, IEEE Trans. Consum. Electron. 69 (3) (2023) 478–489.

[35] B. Yang, G. Bender, Q.V. Le, J. Ngiam, CondConv: conditionally parameterized convolutions for efficient inference, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019, pp. 1307–1318.

[36] X. Li, W. Wang, X. Hu, J. Yang, Selective kernel networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 510–519.

[37] J.-J. Liu, Q. Hou, M.-M. Cheng, C. Wang, J. Feng, Improving convolutional networks with self-calibrated convolutions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10096–10105.

[38] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, et al., Resnest: Split-attention networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2736–2746.

[39] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, X. Li, Large selective kernel network for remote sensing object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 16794–16805.

[40] M. Chen, Q. Zhang, Q. Song, X. Qian, R. Guo, M. Wang, D. Chen, Neural-free attention for monaural speech enhancement toward voice user interface for consumer electronics, IEEE Trans. Consum. Electron. 69 (4) (2023) 765–774.

[41] C. Chen, H. Lu, H. Hong, H. Wang, S. Wan, Deep self-supervised graph attention convolution autoencoder for networks clustering, IEEE Trans. Consum. Electron. 69 (4) (2023) 974–983.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 6000–6010.

[43] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[44] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11534–11542.

[45] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, J. Shlens, Scaling local self-attention for parameter efficient visual backbones, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12894–12904.

[46] Y. Su, H. Xia, Q. Liang, W. Nie, Exposing deepfake videos using attention based convolutional lstm network, Neural Process. Lett. 53 (2021) 4159–4175.

[47] W. Lu, L. Liu, B. Zhang, J. Luo, X. Zhao, Y. Zhou, J. Huang, Detection of deepfake videos using long-distance attention, IEEE Trans. Neural Networks Learn. Syst. 35 (7) (2024) 9366–9379.

[48] H.-Y. Zhou, C. Lu, S. Yang, X. Han, Y. Yu, Preservational learning improves self-supervised medical image models by reconstructing diverse contexts, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3499–3509.

[49] Y. Li, X. Yang, P. Sun, H. Qi, 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.

[50] B. Zi, M. Chang, J. Chen, X. Ma, Y.-G. Jiang, Wilddeepfake: A challenging real-world dataset for deepfake detection, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2382–2390.

[51] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, Retinaface: Single-shot multi-level face localisation in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5203–5212.

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

[53] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, 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, 2021, pp. 772–781.

[54] C. Wang, W. Deng, Representative forgery mining for fake face detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14923–14932.

[55] J. Li, H. Xie, J. Li, Z. Wang, Y. Zhang, Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6458–6467.

[56] H.H. Nguyen, F. Fang, J. Yamagishi, I. Echizen, Multi-task learning for detecting and segmenting manipulated facial images and videos, in: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems, BTAS, IEEE, 2019, pp. 1–8.

[57] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, 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.

[58] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.
[59] I. Masi, A. Killekar, R.M. Mascarenhas, S.P. Gurudatt, W. AbdAlmageed, Two-branch recurrent network for isolating deepfakes in videos, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16, Springer, 2020, pp. 667–684.
[60] D. Zhang, J. Chen, X. Liao, F. Li, J. Chen, G. Yang, Face forgery detection via multi-feature fusion and local enhancement, IEEE Trans. Circuits Syst. Video Technol. 34 (9) (2024) 8972–8977.
[61] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.
[62] Y. Qian, G. Yin, L. Sheng, Z. Chen, J. Shao, Thinking in frequency: Face forgery detection by mining frequency-aware clues, in: European Conference on Computer Vision, Springer, 2020, pp. 86–103.
[63] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
[64] A. Haliassos, K. Vougioukas, S. Petridis, M. Pantic, Lips don't lie: A generalisable and robust approach to face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5039–5049.
[65] L. Jiang, R. Li, W. Wu, C. Qian, C.C. Loy, Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2889–2898.