Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.063141

ARTICLE

SMNDNet for Multiple Types of Deepfake Image Detection

Qin Wang[1], Xiaofeng Wang[2,*], Jianghua Li[2], Ruidong Han[2], Zinian Liu[1] and Mingtao Guo[3]

[1]Department of Computer Science and Engineering, Xi'an University of Technology, Xi'an, 710048, China
[2]Department of Mathematics, Xi'an University of Technology, Xi'an, 710054, China
[3]The National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu, 610065, China

*Corresponding Author: Xiaofeng Wang. Email: xfwang66@sina.com.cn

Received: 06 January 2025; Accepted: 27 February 2025; Published: 19 May 2025

ABSTRACT: The majority of current deepfake detection methods are constrained to identifying one or two specific types of counterfeit images, which limits their ability to keep pace with the rapid advancements in deepfake technology. Therefore, in this study, we propose a novel algorithm, Stereo Mixture Density Network (SMNDNet), which can detect multiple types of deepfake face manipulations using a single network framework. SMNDNet is an end-to-end CNN-based network specially designed for detecting various manipulation types of deepfake face images. First, we design a Subtle Distinguishable Feature Enhancement Module to emphasize the differentiation between authentic and forged features. Second, we introduce a Multi-Scale Forged Region Adaptive Module that dynamically adapts to extract forged features from images of varying synthesis scales. Third, we integrate a Nonlinear Expression Capability Enhancement Module to augment the model's capacity for capturing intricate nonlinear patterns across various types of deepfakes. Collectively, these modules empower our model to efficiently extract forgery features from diverse manipulation types, ensuring a more satisfactory performance in multiple-types deepfake detection. Experiments show that the proposed method outperforms alternative approaches in detection accuracy and AUC across all four types of deepfake images. It also demonstrates strong generalization on cross-dataset and cross-type detection, along with robust performance against post-processing manipulations.

KEYWORDS: Convolutional neural network; deepfake detection; generative adversarial network; feature enhancement

1 Introduction

Deepfake generally uses deep learning methods to mimic the distribution of genuine faces to generate fake face images or videos, which are difficult for human eyes and traditional forensic methods to recognize [1]. With the development of deepfake technology, it has the potential to bring new life to movies, music, video games, and advertisements. However, malicious applications of deepfake technology can have serious consequences [2]. In particular, the release of deepfake speech videos by a state leader can have a damaging impact and pose a threat to national security and counter-terrorism efforts [3]. For example, a video of Obama's public rebuke of Trump has received significant attention, but it was generated by transferring the actor's expression to Obama's face and matching it with a synthetic voice. Moreover, deepfake technology is widely abused in pornographic videos. According to a recent report, over 96% of deepfake content is pornographic. The faces of popular female celebrities are often inserted into pornographic videos, exposing them to serious reputational risks and credibility crises [4]. Furthermore, with the public availability of deepfake technology and the accessibility of personal information on the internet, everyone's images and

videos can be manipulated and used for parody, fraud, and other malicious activities, which could seriously impact personal reputation and financial security. Therefore, the detection of deepfake face images and videos has become an urgent and critical task [5].

At present, deepfake face images/videos mainly fall into four types, entire face synthesis, identity swap/face swap, expression swap, and attribute manipulation [3]. For each deepfake type, a variety of deepfake detection methods have been developed. However, these strategies can only detect certain deepfake types of images or videos, which merely need to learn the single, specific forgery artifact features. With the development of deepfake detection technology, some methods capable of detecting both face swap and expression swap simultaneously have been gradually proposed, since the two types of forged videos involve local forgery, which has a certain uniformity. However, with the gradual maturity and widespread application of Generative Adversarial Networks (GANs), various forged face imageries are inundated with the social network, previous detection methods that can only detect one or two types of forgeries are inefficient, i.e., they lack versatility.

The versatility of the detection methods is very important because multiple types of forgery need to be distinguished in many applications such as criminal investigation and forensic evidence. Therefore, there is an urgent need to develop detection methods that can detect multiple types of forgeries. In this study, we consider the four primary deepfake manipulations as the detection objects and propose a new method to detect them using a single network, SMNDNet. The algorithm flow of the proposed method is shown in Fig. 1. First, we preprocess the existing identity swap and expression swap videos to generate corresponding image data. Second, we utilize some existing and self-generated entire face synthesis and attribute manipulation images, and generated identity swap and expression swap images to compose the Large-scale Deepfake Face Image (LDFI) database. Third, for the feature distribution of four different deepfake and real images, we use SMNDNet to extract the distinguishable features and perform classification detection on the extracted features. In the feature extraction phase, we provide a Subtle Distinguishable Feature Enhancement (SDFE) module to enhance the subtle feature differences between genuine and fake face images. Additionally, we introduce the Multi-scale Forged Region Adaptive (MFRA) module and the Nonlinear Expression Capability Enhancement (NECE) module to improve the network's adaptability to all types of deepfake face images. These two modules can respectively adjust to different forging ranges and various generation modes of different forging types. In the detection phase, the detection module classifies the extracted features into two categories. Experimental results show that our method has satisfactory classification accuracy and adaptability to various forgery types.

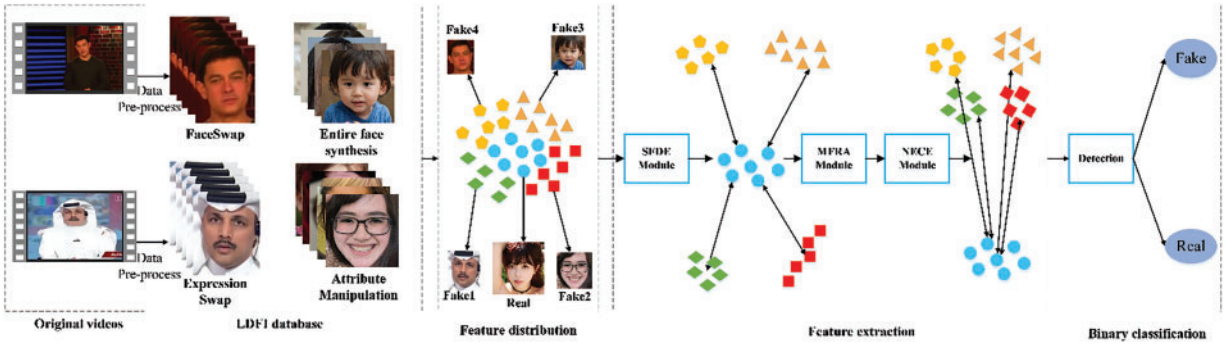


Figure 1: The overall workflow of the proposed method

The contributions of this study are:

(1) Specifically for detecting multi-forgery types of deepfake images, we propose an end-to-end CNN-based network, SMNDNet, that consists of the SDFE module, MFRA module, NECE module, and detection module.
(2) Using the proposed network model, we develop a deepfake face image detection method that can detect entire face synthesis, identity swap, expression swap, and attribute manipulation. Experimental results show that the method achieves satisfactory detection accuracy, generalization, and robustness.
(3) We collect a Large-scale Deepfake Face Image (LDFI) database that consists of the entire face synthesis dataset, identity swap dataset, expression swap dataset, and attribute manipulation dataset.

2 Related Works

In this section, we summarize the existing deepfake face imagery detection methods and classify them into three categories. The first category is the methods that can only detect one type of forgery, and the methods are classified in detail according to the type they can detect. The second category is the methods that can simultaneously detect two types of forgery (face swap and expression swap). The third category shows the multi-type deepfake detection methods that can detect four types of forgery.

Early detection methods mostly focused on a single type of forgery that only distinguishes one manipulation type such as entire face synthesis, face swap, expression swap, or attribute manipulation from genuine camera imagery. For detecting the entire face synthesis images, Dang et al. [6] proposed a customized CNN network model IF-CGFace. Some scholars [7–9] proposed to detect deepfakes by extracting GAN fingerprints from images. Hsu et al. [10] proposed a two-step learning DeepFD network. For detecting attribute manipulation images, in [11], Marra et al. presented a detection method using a GAN fingerprint. Marra et al. [12] proposed an incremental learning method. Nataraj et al. [13] proposed a CNN-based method that uses the gray-level co-occurrence matrix. For detecting face swap images, Zhou et al. [14] proposed a two-stream network that applies the low-level camera features and local noise residual features. Heo et al. [15] presented an improved vision transformer model. Guo et al. [16] proposed an adaptive fusion-based guided residuals network (AdapGRnet). Yang et al. [17] proposed an SA-DTH-net (Speaker Authentication network based on Dynamic Talking Habits) to detect expression swap videos. Generally, these methods can obtain satisfactory detection results on the dataset generated by specific GANs, however, with the new GANs emerging constantly, they can only distinguish a specific forgery type from real imagery, and the versatility is not satisfactory.

With the development of artificial intelligence technology, the functionality of GAN is becoming increasingly powerful. In recent years, faceswap and expression swap have become the main types of deepfake and caused serious adverse impacts. Therefore, some excellent detection methods for both face swap and expression swap manipulations have been presented. Shiohara et al. [18] exploited the affine artifact features to facilitate forgery image detection. Qiao et al. [19] proposed an unsupervised Deepfake detector that utilizes an enhanced contrastive learning method. Sharma et al. [20] introduced a GAN-CNN to reduce catastrophic forgetting, improving detection accuracy. Sharma et al. [21] utilized a Compact Ensemble-based discriminators framework integrated with Deep Conditional Generative Adversarial Networks (CED-DCGAN) for identifying real-time deepfakes. Peng et al. [22] proposed a Deepfake gaze analysis (DFGaze) approach that identifies spatial-temporal inconsistencies in gaze features within deepfake videos. Yu et al. [23] utilized multi-task learning methodologies to detect deepfakes. Song et al. [24] proposed incorporating the concept of sample hardness into the training process of deepfake detectors through a curriculum learning paradigm. Guo et al. [25] proposed a groundbreaking Space-Frequency Interactive Convolution (SFIConv) technique. Wang et al. [26] proposed a cutting-edge Complementary Dynamic Interaction Network (CDIN). Zhao

et al. [27] proposed a new framework, TAN-GFD, to extract texture-based information and adaptive noise mining. All the above methods can detect face swap and expression swap manipulations simultaneously, but they cannot be extended to detect other types of forgery.

As far as we know, there are only two works that can detect the four main types of forgery. Wang et al. [28] proposed an algorithm FakeSpotter that detects deepfake face images by monitoring the behavior of neurons. In this method, the authors used the number of activated neurons as the discriminative feature, and input it into the fully connected network for binary classification. Dang et al. [29] proposed to use of an attention mechanism to emphasize the identification features of forged regions and input the features into CNN for classification. This method not only can detect the deepfake image, but also locate the manipulated region.

However, the two methods used a fully connected network and an existing CNN-based network as classifiers. Therefore, it is of great significance to propose a well-designed CNN-based network specialized in detecting multiple types of deepfakes.

3 The Proposed Method

In this section, we propose a novel deepfake face detection algorithm that can detect four main deepfake types using a single well-designed network model, SMNDNet. The framework of SMNDNet is shown in Fig. 2. First, we analyze the design motivation of the SMNDNet. Second, we describe the network structure. Third, we describe the loss function and optimizer.

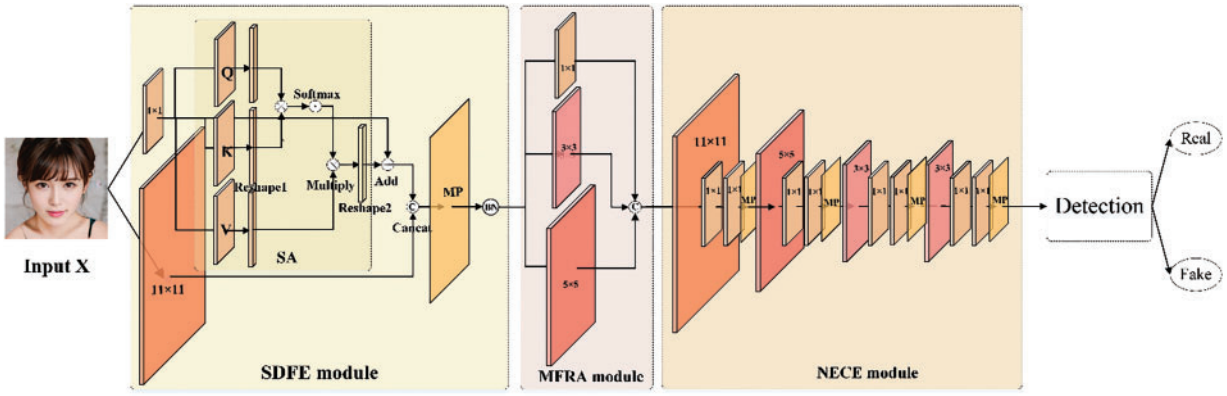


Figure 2: The framework of SMNDNet

3.1 Design Motivation of the SMNDNet

In order to detect multiple types of deepfakes, first, it is necessary to extract distinguishable features between various types of deepfake face images and genuine face images. Second, it needs to classify the extracted distinguishable features. Therefore, we customized a novel network model that consists of a Subtle distinguishable feature enhancement module, a Multi-scale forged region adaptive module, a Nonlinear expression capability enhancement module, and a Detection module. The first three modules are used for feature extraction and the last module is used for classification. We named our network SMNDNet, using the initials of the four modules.

To extract distinguishable features between various types of deepfake face images and genuine face images, we deeply analyze three characteristics of deepfake face images and design three modules based on them. First, considering that the deepfake face image is completely simulated the genuine face image distribution generated, almost cannot be distinguished by human eyes, only subtle distinguishable features

concentrate on the facial features and the facial contour area. To capture the distinguishable features between genuine and pseudo images, we provide a Subtle Distinguishable Feature Enhancement (SDFE) module to enhance the facial features and facial contours, further guiding the network to focus more on the discrepancy regions between genuine and fake images. Second, by analyzing the generation process of various pseudo-face images, we discovered that their modified areas were different. While the alternative pixels of face swap images only in the facial region and the expression swap manipulation merely altered the pixel distribution of the identity places (such as the eyes, mouth, etc.) in the original genuine face image, the whole image created by the entire face synthesis and the attribute manipulation was entirely synthesized. To use a single network to efficiently capture the forged features of various forgeries, we propose a Multi-scale Forged Region Adaptive (MFRA) module, which enables the network to adaptively extract forged features of different forgery sizes. Third, although the GAN approach is utilized for entire face synthesis, attribute manipulation, and faceswap, which has a completely distinct generation process, and the expression swap is done via a graphical method. Hence, the distribution of the four types of forged images is entirely different. To adapt to four different pixel distributions, the network requires extremely strong nonlinear expression capabilities. For this reason, we propose a Nonlinear Expression Capability Enhancement (NECE) module, in which we develop four similar nonlinear blocks. To sum up, the SDFE module enhances the discrepancy features between genuine and fake images. The MFRA module and NECE module extract nonuniformity among various forged images.

To classify the extracted distinguishable features, we design a detection module that includes a global average pooling layer, two fully connected layers, and a Dropout layer. For the classification task, the fully connected operation is usually used to classify the result of feature extraction. However, only using the fully connected operation will greatly increase the number of parameters of the network model, thus the global average pooling (GAP) is used to reduce the number of parameters and obtain global feature information. Considering two fully connected layers contain a large number of parameters, which would increase the training time greatly and cause over-fitting. Therefore, the Dropout mechanism is added between two fully connected layers and the dropout rate is set to 0.5. By ignoring half of the neurons in each training batch, training time and over-fitting can be significantly reduced.

3.2 Network Structure of SMNDNet

In this section, we describe the network structure of the proposed method. The proposed network model includes four modules: SDFE module, MFRA module, NECE module, and Detection module.

Subtle Distinguishable Feature Enhancement Module

Given an input RGB image X, The SDFE module takes X as input and generates the enhanced subtle distinguishable feature S.

In the SDFE module, firstly, we propose a two-stream structure that applies regular small 1×1 convolution and large 11×11 convolution on X to procure X_s and X_l, respectively. The small convolution is used to keep the image's spatial correlation and align the channel dimension of the following operation. The large convolution is used to capture the image's large-scale correlation features to improve the performance of the next task. Following the small convolution, we next apply the Self-Attention (SA) block to get the feature map by adding greater weight to the positions of discernible features in the image. After that, we concatenate the output features of the two streams to acquire the feature map, seeing in Eq. (1).

$$X_c = Concat\left(X_{SA}, X_l\right) \tag{1}$$

where $Concat\left(.\right)$ dedicates the concatenation operation of feature maps. The X_{SA} is generated by the SA block.

In SA block, takes X_s as input, three 1×1 convolutions are executed on it to obtain X_Q, X_K, and X_V firstly, then the $R1(.)$ is used to resize three-dimensional vectors $X_Q \in R^{H \times W \times C}$, $X_K \in R^{H \times W \times C}$, and $X_V \in R^{H \times W \times C}$ into two-dimensional vectors $X_q \in R^{N \times C}$, $X_k \in R^{N \times C}$, $X_v \in R^{N \times C}$, $N = H \times W$, seeing in Eq. (2).

$$X_q = R1\left(X_Q\right), X_k = R1(X_K), X_v = R1(X_V) \tag{2}$$

Here, $R1(.)$ is the resize operation that transforms the input tensor X from a three-dimensional shape (H, W, C) to a two-dimensional shape (H × W, C). Specifically, this operation flattens the heigh H and width W dimensions into a single dimension while maintaining the number of channels C at each position. Conceptually, it can be viewed as sequentially expanding the H × W elements into one dimension in row-major order, thereby producing a new tensor with the shape (H × W, C). After that, we multiply X_q and X_k, where the multiplication of the query (X_q) and key (X_k) is a core operation. This multiplication computes the similarity or relevance between the query and key, indicating which keys the query should focus on. After this multiplication, we apply the $Softmax(.)$ to normalize the scores, resulting in a probability distribution that represents the attention weights of each query with respect to the keys. This process ensures that the attention scores effectively influence the weighted sum, allowing the model to accurately weigh the importance of each position when computing the value (X_v). Then, the $R2(.)$ is used to resize the two-dimensional vector to three-dimensional vector $X_M \in R^{H \times W \times C}$, seeing in Eq. (3).

$$X_M = R2\left(X_v \otimes Softmax\left(X_q \otimes X_k\right)\right) \tag{3}$$

In the end, the output X_{SA} of the SA block can be obtained, using Eq. (4).

$$X_{SA} = \lambda \times X_M + X_s \tag{4}$$

where λ is a learnable scalar initialized to 0. Initializing λ to 0 provides a stable starting point for training, enabling the model to progressively learn the optimal balance between the input X_s and the outputs X_M. This initialization strategy enhances both training stability and flexibility.

Following the SA block, the MaxPooling (MP) layer and BatchNormalization (BN) layer are connected to obtain the SDFE block's output feature map S, seeing Eq. (5).

$$S = BN\left(MP\left(X_c\right)\right) \tag{5}$$

Multi-Scale Forged Region Adaptive Module

In this module, we design convolution kerns of scale 1, 3 and 5, in which large convolution kerns could adaptively extract long distance dependent features, while tiny convolution kerns could extract short distance dependent features. The MFRA module is defined as Eq. (6).

$$M = Concat\left(conv_{1\times1}\left(S\right), conv_{3\times3}\left(S\right), conv_{5\times5}\left(S\right)\right) \tag{6}$$

where M is the output map and S is the module input. The $conv_{i\times i}(.)$ dedicates the convolution operation with the kernel size $i \times i$.

Nonlinear Expression Capability Enhancement Module

In the NECE module, we develop four similar nonlinear blocks. Each nonlinear block includes three convolution layers and one MaxPooling layer, except for the first convolution layer, the kernels of the other layers have the same size. Given an input M, the NECE module takes M as input and generates N_1, N_2, N_3,

and N_4 in each nonlinear block. The outputs of the first convolution layer in each nonlinear block are N_1^1, N_2^1, N_3^1, and N_4^1, respectively, seeing Eq. (7).

$$\begin{cases} N_1^1 = conv_{11\times11}(M) \\ N_2^1 = conv_{5\times5}(N_1) \\ N_3^1 = conv_{3\times3}(N_2) \\ N_4^1 = conv_{3\times3}(N_3) \end{cases} \tag{7}$$

In each nonlinear block, the outputs of the second convolution layer are N_j^2, the outputs of the third convolution layer are N_j^3, here, $j = 1, 2, 3, 4$, seeing Eq. (8).

$$\begin{cases} N_j^2 = ReLU(conv_{1\times1}(N_j^1)) \\ N_j^3 = ReLU(conv_{1\times1}(N_j^2)) \\ N_j = MaxPooling(N_j^3) \end{cases} \tag{8}$$

The activation function $ReLU(.)$ is shown in Eq. (9).

$$ReLU(x) = \begin{cases} x, x > 0 \\ 0, 0 \le 0 \end{cases} \tag{9}$$

Detection Module

The detection module takes N_4 as input and output feature D. The detection module is calculated by Eq. (10).

$$D = FC_2(Dropout_{0.5}(FC_{1024}(GAP(N_4)))) \tag{10}$$

where the $GAP(.)$ is the GAP operation, the $FC_i(.)$ dedicates the FC operation and i is the output neuron numbers. The $Dropout_{0.5}(.)$ is the Dropout operation.

3.3 Loss Function and Optimizer

Loss function: In the training stage, the cross-entropy loss function is used to observe the change in loss value. The cross-entropy loss function in Eq. (11).

$$Loss = -\sum_{i=0}^{1} Y_i \cdot logD_i \tag{11}$$

where i represents class label, Y_i represents the real/pseudo label of the input image ($i = 0$ or 1), D_i represents the output of the input image on the SMNDNet.

Optimizer: we use the stochastic gradient descent (SGD) optimization algorithm to optimize the loss function $Loss$. SDG updates the parameter θ as follows:

$$\theta = \theta - \alpha \frac{\partial}{\partial \theta} Loss \tag{12}$$

where α the learning rate.

4 Experimental Results and Analysis

In this section, we evaluate the performance of the proposed method through a series of experiments. The simulation settings of the proposed method are shown in Table 1.

Table 1: Simulation settings

Setting	Description
Operating system	Windows 10
CPU	Intel i7-6800K
GPU	NVIDIA GTX 1080Ti
RAM	32 GB
Input image size	224 × 224 × 3
Optimizer	Stochastic Gradient Descent (SGD)
Learning rate	0.01
Batch size	4
Number of epochs	100
Evaluation metrics	Precision (Pre.), Accuracy (Acc.), Recall (Rec.), AUROC/AUC

4.1 Dataset

To detect four types of deepfake manipulations, we present the LDFI dataset, as detailed in Table 2. The LDFI encompasses four types of datasets: Entire Face Synthesis Fake (EFSF), Identity Swap Fake Face (ISFF), Expression Swap Fake Face (ESFF), and Attribute Manipulation Fake Face (AMFF). Notably, the EFSF includes two sub-datasets: high-definition $EFSF_1$ and standard-definition $EFSF_2$. The sources for these datasets are diverse, comprising three primary origins: existing datasets (existing), self-generated data (self-generated), and video datasets from which face regions were extracted using the MTCNN algorithm [30] (extracted). Each dataset is partitioned into training, validation, and testing sets, with the table indicating the counts of both real and fake samples for each set. This systematic approach ensures a comprehensive evaluation across various deepfake manipulation types, providing a robust foundation for assessing the proposed method's performance.

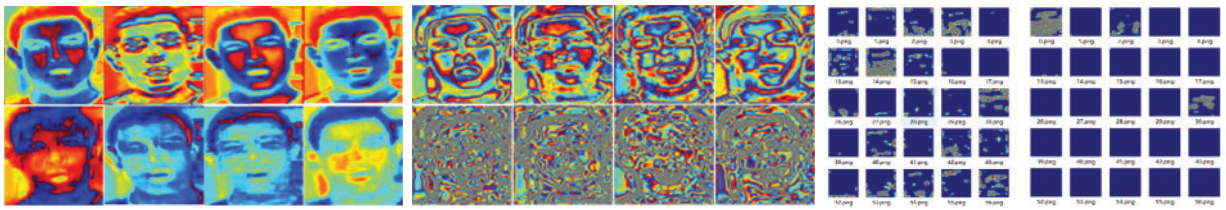
Table 2: LDFI database

Name	Data	Origins	Training	Validation	Testing
$EFSF_1$	StyleGAN2 [31]	Existing	8000	1000	1000
$EFSF_2$	DCGAN [32]	Self-generated	8000	1000	1000
	PGGAN [29]	Existing	8000	1000	1000
	StyleGAN [29]	Existing	8000	1000	1000
ISFF	Celeb_DF [33]	Extracted	8000	1000	1000
ESFF	NeuralTexture [34]	Extracted	8000	1000	1000
	Face2Face [34]	Extracted	8000	1000	1000
AMFF	FaceAPP [29]	Existing	8000	1000	1000
	StarGAN [29]	Existing	8000	1000	1000

4.2 Ablation Study

To demonstrate the rationality of the model design, we conduct ablation studies by replacing specific modules with convolutional layers of varying sizes. Specifically, we substitute the SDFE module with an 11 × 11 convolution, the MFRA module with a 3 × 3 convolution, and modify the NECE module using four convolutions. For the NECE module modification, we replace each complete nonlinear block with its first convolution layer. All ablation experiments are trained on the ISFF dataset.

Firstly, we visualize the output feature maps of each module (SDFE, MFRA, NECE) and those obtained using the replacement modules, as illustrated in Fig. 3. In Fig. 3a,b, the first row shows the feature maps with the corresponding module, the second row shows feature maps without corresponding module. In Fig. 3c, the left shows the partial channel of feature maps N4 in our network, the right shows feature maps without the NECE module. We observe that from Fig. 3a, compared to standard convolution outputs, the SDFE module's feature maps exhibit an enhanced emphasis on facial features and contours. This indicates the SDFE module's capability to direct the network's attention toward regions with subtle yet distinguishable characteristics. As shown in Fig. 3b, the incorporation of the MFRA module allows the network to capture clearer and more nuanced features relative to standard convolution. From Fig. 3c, it is evident that the NECE module generates a significantly higher number of channels with nonlinear features compared to standard convolution. Consequently, the introduction of the NECE module endows the network with superior nonlinear expression capabilities, enabling it to perform more complex tasks.



(**a**) SDFE module (**b**) MFRA module (**c**) NECE module

Figure 3: The visual results of the ablation experiment

Secondly, we evaluate the detection performance of the model with all modules intact vs. the performance when any single module (such as SDFE, MFRA, or NECE) is removed, as detailed in Table 3. In the table, bold text indicates the maximum value. As evidenced by Table 3, the removal of any individual module adversely impacts the overall detection performance, indicating that the incorporation of the SDFE, MFRA, and NECE modules is essential for achieving optimal results. This finding further substantiates the rationale behind our model design.

Table 3: The numerical results of the ablation experiment

	No SDFE	No MFRA	No NECE	Original model
Acc. (%)	95.53	97.73	95.93	**98.74**
AUC (%)	96.59	98.76	96.56	**99.93**

4.3 Analysis and Comparison of the Detection Performance

We test the detection performance of the proposed method and compare it with the other methods which are classical network frameworks and state-of-the-art deepfake detection methods.

4.3.1 Comparison with Classic Network Frameworks

We apply dataset $EFSF_1$ which provides with the clearest and most realistic deepfake face images to test the proposed SMNDNet and the compared networks, ResNet50 [35], ResNet18 [35], Inceptionv3 [36], and Xception [37]. For the fairness of the experimental comparison, we set the same experimental settings for these comparison networks, the test results are shown in Table 4. The bold font represents the best results on this evaluation metric. As can be seen from Table 4, the evaluation metrics of our method are higher than those of four existing classical networks. Moreover, the number of parameters in our method is significantly less than that of all the other classic network models. In summary, the proposed method stands out for its lightweight network framework but yields satisfactory deepfake face image detection.

Table 4: Compare with four classical CNN-based networks

CNN Network	Acc. (%)	Pre. (%)	Rec. (%)	AUC (%)	Number of parameters
ResNet50 [35]	82.63	99.89	65.33	98.76	23,591,810
ResNet18 [35]	98.00	98.56	97.42	99.84	12,568,194
Inceptionv3 [36]	98.03	98.56	97.49	99.77	21,806,882
Xception [37]	98.03	99.43	96.62	99.93	20,865,578
SMNDNet	**99.07**	**99.93**	**98.22**	**99.97**	**10,812,603**

4.3.2 Comparison with the State-of-the-Art Deepfake Detection Methods

To further substantiate the efficacy of our proposed method, we conducted a comparative analysis against both single-type and multi-type deepfake detection methods, as detailed in Tables 5 and 6. In Tables 5 and 6, $EFSF_2$ serves as the training dataset for the entire face synthesis forgery method, while the ISFF, ESFF, and AMFF datasets are utilized to train the respective manipulation type methods. And E.F.S., A.M., E.S., and F.S. denote entire face synthesis, attribute manipulation, expression swap, and face swap, respectively. As shown in Table 5, our model outperforms all single-type detection methods on corresponding manipulation types. Furthermore, Table 6 illustrates that our method exhibits enhanced detection results compared to another multi-types detection method on four manipulation types. In Tables 5 and 6, bold text indicates the maximum value.

4.4 Generalization

To evaluate the generalization capability of the proposed method, we employ distinct datasets for the training and testing phases. In pursuit of comprehensive experimental coverage, we perform two sets of experiments: one emphasizing cross-dataset analysis and the other focusing on cross-type analysis. The results are summarized in Table 7.

As shown in Table 7, our method achieves a detection accuracy and AUC value exceeding 80% in both experiments. This indicates that the algorithm demonstrates consistent generalization performance on both cross-dataset and cross-type detection.

Table 5: Comparison with single-type deepfake detection methods: AUC (%)

Entire face synthesis		Attribute manipulation		Faceswap		Expression swap	
Methods	**AUC**	**Methods**	**AUC**	**Methods**	**AUC**	**Methods**	**AUC**
Barni et al. [38]	98.62	Nataraj et al. [13]	96.46	Zhou et al. [14]	92.70	Chen et al. [39]	99.05
Dang et al. [6]	99.01	Marra1 et al. [12]	99.37	Li et al. [40]	95.93	Afchar et al. [41]	94.60
Albright et al. [8]	97.89	Marra2 et al. [11]	99.20	Guo et al. [16]	99.70	Zhao et al. [27]	98.18
McCloskey et al. [42]	92.00	–	–	Agarwal et al. [43]	98.50	Shiohara et al. [18]	99.88
Yu et al. [7]	96.23	–	–	Heo et al. [15]	98.59	Guo et al. [16]	98.36
Guarnera et al. [9]	97.56	–	–	Guo et al. [25]	99.01	Wang et al. [26]	98.69
Ours	**99.99**	Ours	**99.60**	Ours	**99.91**	Ours	**99.97**

Table 6: Comparison with multiple-types deepfake detection method: FakeSpotter (FS)

	Pre.		Rec.		Acc.		AUC	
	FS [28]	**Ours**	**FS**	**Ours**	**FS**	**Ours**	**FS**	**Ours**
E. F. S.	94.94	**97.47**	95.68	**97.22**	95.33	**97.36**	95.26	**99.41**
A. M.	89.36	**99.97**	89.22	**99.98**	89.12	**99.95**	89.27	**99.91**
E. S.	94.91	**97.93**	94.84	**98.42**	74.63	**94.95**	94.94	**95.90**
F. S.	83.55	**98.78**	85.63	**98.50**	83.42	**98.67**	83.39	**99.94**

Table 7: Cross-dataset and cross-type deepfake detection results

Cross-dataset				Cross-type			
Train	**Test**	**Acc. (%)**	**AUC (%)**	**Train**	**Test**	**Acc. (%)**	**AUC (%)**
$EFSF_1$	$EFSF_2$	84.00	85.14	$EFSF_{1,2}$	AMFF: FaceAPP	99.18	99.95
$EFSF_2$	$EFSF_1$	80.72	81.45	$EFSF_{1,2}$	AMFF: StarGAN	97.73	99.87

4.5 Robustness Analysis

To evaluate the robustness of the proposed method, we systematically investigate its detection performance under various post-processing manipulations. These manipulations encompass Rotation (angle: 5 degrees), Shear (range: 0.2), Zoom (range: 0.2), Height shift (rate: 0.5), Width shift (rate: 0.2), Flip, and Zero-phase Component Analysis (ZCA) whitening. We utilize the $EFSF_1$ dataset as the training set, apply each manipulation individually to $EFSF_1$, and generate corresponding test datasets. The results are summarized in Table 8. In this table, bold text denotes the average maximum value and original value, while italic text indicates the average minimum value. The 'Original' condition refers to the absence of any post-processing manipulations on the test set.

As illustrated in Table 8, despite the application of Height shift manipulation on the test images, the AUC decreases by 7.01% compared to scenarios without any post-processing manipulations. However, all other operations reduce AUC by no more than 3%. Additionally, when Shear and ZCA_whitening manipulations are applied to the test images, the experimental results remain nearly identical to those obtained without post-processing manipulations. In summary, our method shows satisfactory robustness against post manipulations.

Table 8: Test results on different post-processing manipulations: AUC

Manipulations	Internet celebrity (%)	Asian star (%)	Yellow race (%)	Average (%)
Original	99.9450	99.9451	99.4479	**99.7793**
Rotation	99.6183	99.6298	97.6640	98.9707
Shear	99.9471	99.9428	99.4346	99.7748
Zoom	98.1945	98.3303	96.3422	97.6223
W_shift	99.8575	99.7797	98.0329	99.2234
H_shift	95.1741	95.5361	87.3269	*92.6790*
Flip	99.9362	99.9660	98.8691	99.5904
ZCA_whitening	99.9450	99.9451	99.4380	**99.7760**

5 Conclusions

In this study, we propose an end-to-end CNN-based network, SMNDNet, for detecting four types of deepfake face images, including entire face synthesis, identity swap, attribute manipulation, and expression swap. Compared with some classic networks such as Xception, and ResNet50, our proposed SMNDNet achieves the best detection results with the minimum number of network parameters. Compared with existing deepfake detection methods, the proposed method has the advantage of using a single network framework to detect multiple types of deepfake manipulations while achieving higher detection accuracy than other methods that can only detect one type of deepfake manipulation. Furthermore, SMNDNet outperforms other multi-type detection methods in terms of detection performance. Additionally, our method exhibits strong robustness and generalization across both cross-dataset and cross-type detection scenarios.

Acknowledgement: The authors would like to express their gratitude for the valuable feedback and suggestions provided by all the anonymous reviewers and the editorial team.

Funding Statement: This research was funded by the National Natural Science Foundation of China (Grant No. 62376212) and the Shaanxi Science Foundation of China (Grant No. 2022GY-087). These grants were awarded to Professor Xiaofeng Wang. This research was also supported by the Open Fund of Intelligent Control Laboratory. The grant was awarded to Professor Jianghua Li.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Qin Wang and Xiaofeng Wang; methodology, Qin Wang and Xiaofeng Wang; software, Qin Wang; validation, Qin Wang, Xiaofeng Wang, Jianghua Li, Ruidong Han, Zinian Liu, and Mingtao Guo; formal analysis, Qin Wang and Ruidong Han; investigation, Qin Wang and Jianghua Li; resources, Xiaofeng Wang; data curation, Mingtao Guo; writing—original draft preparation, Qin Wang; writing—review and editing, Xiaofeng Wang; visualization, Mingtao Guo; supervision, Xiaofeng Wang; project administration, Xiaofeng Wang; funding acquisition, Xiaofeng Wang and Jianghua Li. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Xiaofeng Wang, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Seow JW, Lim MK, Phan RCW, Liu JK. A comprehensive overview of Deepfake: generation, detection, datasets, and opportunities. Neurocomputing. 2022;513(1):351–71. doi:10.1016/j.neucom.2022.09.135.
2. Westerlund M. The emergence of deepfake technology: a review. Technol Innov Manag Rev. 2019;9(11):39–52. doi:10.22215/timreview/1282.
3. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J. Deepfakes and beyond: a survey of face manipulation and fake detection. Inf Fusion. 2020;64(1):131–48. doi:10.1016/j.inffus.2020.06.014.
4. Lyu S. Deepfake detection: current challenges and next steps. In: 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW); 2020 Jul 6–10; London, UK. p. 1–6.
5. Yu P, Xia Z, Fei J, Lu Y. A survey on deepfake video detection. IET Biom. 2021;10(6):607–24. doi:10.1049/bme2.12031.
6. Dang LM, Hassan SI, Im S, Lee J, Lee S, Moon H. Deep learning based computer-generated face identification using convolutional neural network. Appl Sci. 2018;8(12):2610. doi:10.3390/app8122610.
7. Yu N, Davis LS, Fritz M. Attributing fake images to GANs: learning and analyzing GAN fingerprints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea; p. 7556–66.
8. Albright M, McCloskey S. Source generator attribution via inversion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 16–20; Long Beach, CA, USA. p. 96–102.
9. Guarnera L, Giudice O, Battiato S. Deepfake detection by analyzing convolutional traces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2020 Jun 14–19; Seattle, WA, USA. p. 666–7.
10. Hsu CC, Zhuang YX, Lee CY. Deep fake image detection based on pairwise learning. Appl Sci. 2020;10(1):370. doi:10.3390/app10010370.
11. Marra F, Gragnaniello D, Verdoliva L, Poggi G. Do GANs leave artificial fingerprints? In: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR); 2019 Mar 28–30; San Jose, CA, USA. p. 506–11.
12. Marra F, Saltori C, Boato G, Verdoliva L. Incremental learning for the detection and classification of GAN-generated images. In: 2019 IEEE International Workshop on Information Forensics and Security (WIFS); 2019 Dec 9–12; Delft, The Netherlands. p. 1–6.
13. Nataraj L, Mohammed T, Manjunath B, Chandrasekaran S, Flenner A, Bappy J, et al. Detecting GAN generated fake images using co-occurrence matrices. Appl Sci. 2019;5(5):532–8. doi:10.2352/ISSN.2470-1173.2019.5.MWSF-532.
14. Zhou P, Han X, Morariu VI, Davis LS. Two-stream neural networks for tampered face detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2017 Jul 21–26; Honolulu, HI, USA. p. 1831–9.
15. Heo YJ, Yeo WH, Kim BG. Deepfake detection algorithm based on improved vision transformer. Appl Intell. 2023;53(7):7512–27. doi:10.1007/s10489-022-03867-9.
16. Guo Z, Yang G, Chen J, Sun X. Exposing deepfake face forgeries with guided residuals. IEEE Trans Multimed. 2023;25:8458–70. doi:10.1109/TMM.2023.3237169.
17. Yang CZ, Ma J, Wang S, Liew AWC. Preventing deepfake attacks on speaker authentication by dynamic lip movement analysis. IEEE Trans Inf Forensics Secur. 2020;16:1841–54. doi:10.1109/TIFS.2020.3045937.
18. Shiohara K, Yamasaki T. Detecting deepfakes with self-blended images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 18720–9.
19. Qiao T, Xie S, Chen Y, Retrain F, Luo X. Fully unsupervised deepfake video detection via enhanced contrastive learning. IEEE Trans Pattern Anal Mach Intell. 2024;46(7):4654–68. doi:10.1109/TPAMI.2024.3356814.
20. Sharma P, Kumar M, Sharma HK. GAN-CNN Ensemble: a robust deepfake detection model of social media images using minimized catastrophic forgetting and generative replay technique. Procedia Comput Sci. 2024;235(1):948–60. doi:10.1016/j.procs.2024.04.090.

21. Sharma SK, AlEnizi A, Kumar M, Alfarraj O, Alowaidi M. Detection of real-time deep fakes and face forgery in video conferencing employing generative adversarial networks. Heliyon. 2024;10(17):e37163. doi:10.1016/j.heliyon.2024.e37163.
22. Peng C, Miao Z, Liu D, Wang N, Hu R, Gao X. Where deepfakes gaze at? Spatial-temporal gaze inconsistency analysis for video face forgery detection. IEEE Trans Inf Forensics Secur. 2024;19:4507–17. doi:10.1109/TIFS.2024.3381823.
23. Yu Z, Cai R, Li Z, Yang W, Shi J, Kot AC. Benchmarking joint face spoofing and forgery detection with visual and physiological cues. IEEE Trans Dependable Secur Comput. 2024;21(5):4327–42. doi:10.1109/TDSC.2024.3352049.
24. Song W, Lin Y, Li B. Towards generic deepfake detection with dynamic curriculum. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2024 Apr 14–19; Seoul, Republic of Korea. p. 4500–4.
25. Guo Z, Jia Z, Wang L, Wang D, Yang G, Kasabov N. Constructing new backbone networks via space-frequency interactive convolution for deepfake detection. IEEE Trans Inf Forensics Secur. 2024;19:401–13. doi:10.1109/TIFS.2023.3324739.
26. Wang H, Liu Z, Wang S. Exploiting complementary dynamic incoherence for deepfake video detection. IEEE Trans Circuits Syst Video Technol. 2023;33(8):4027–40. doi:10.1109/TCSVT.2023.3238517.
27. Zhao Y, Jin X, Gao S, Wu L, Yao S, Jiang Q. TAN-GFD: generalizing face forgery detection based on texture information and adaptive noise mining. Appl Intell. 2023;53(16):19007–27. doi:10.1007/s10489-023-04462-2.
28. Wang R, Ma L, Juefei-Xu F, Xie X, Wang J, Liu Y. FakeSpotter: a simple baseline for spotting AI-synthesized fake faces. arXiv:1909.06122. 2019.
29. Dang H, Liu F, Stehouwer J, Liu X, Jain AK. On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 5781–90.
30. Zhang K, Zhang Z, Li Z, Qiao Y. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett. 2016;23(10):1499–503. doi:10.1109/LSP.2016.2603342.
31. Face website A. AI face website. [cited 2025 Jan 1]. Available from: http://www.seeprettyface.com/mydataset_page2.html#dataset2.
32. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434. 2015.
33. Li Y, Yang X, Sun P, Qi H, Lyu S. Celeb-DF: a large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 3207–16.
34. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. FaceForensics: a large-scale video dataset for forgery detection in human faces. arXiv:1803.09179. 2018.
35. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8.
36. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 2818–26.
37. Chollet F. Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA. p. 1251–8.
38. Barni M, Kallas K, Nowroozi E, Tondi B. CNN detection of GAN-generated face images based on cross-band co-occurrences analysis. In: 2020 IEEE International Workshop on Information Forensics and Security (WIFS); 2020 Dec 4–11; New York, NY, USA. p. 1–6.
39. Chen H, Lin Y, Li B, Tan S. Learning features of intra-consistency and inter-diversity: keys towards generalizable deepfake detection. IEEE Trans Circuits Syst Video Technol. 2022;33(3):1468–80. doi:10.1109/TCSVT.2022.3209336.
40. Li Y, Lyu S. Exposing deepfake videos by detecting face warping artifacts. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2019 Jun 16–17; Long Beach, CA, USA. p. 46–52.

41. Afchar D, Nozick V, Yamagishi J, Echizen I. Mesonet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS); 2018 Dec 11–13; Hong Kong, China. p. 1–7.
42. McCloskey S, Albright M. Detecting GAN-generated imagery using saturation cues. In: 2019 IEEE International Conference on Image Processing (ICIP); 2019 Sep 22–25; Taipei, Taiwan. p. 4584–8.
43. Agarwal S, Farid H, El-Gaaly T, Lim SN. Detecting deep-fake videos from appearance and behavior. In: 2020 IEEE International Workshop on Information Forensics and Security (WIFS); 2020 Dec 6–11; New York, NY, USA. p. 1–6.