MTD-Net: Learning to Detect Deepfakes Images by Multi-Scale Texture Difference

Jiachen Yang, *Member, IEEE*, Aiyun Li, Shuai Xiao, Wen Lu, *Member, IEEE*, and Xinbo Gao, *Senior Member, IEEE*

***Abstract*—With the rapid development of face manipulation technology, it is difficult for human eyes to distinguish fake face images. On the contrary, Convolutional Neural Network (CNN) discriminators can quickly reach high accuracy in identifying fake/real face images. In this study, we explore the behavior of CNN models in distinguish fake/real faces. We find multi-scale texture difference information plays an important role in face forgery detection. Motivated by the above observation, we propose a new Multi-scale Texture Difference model coined as MTD-Net for robust face forgery detection, which leverages central difference convolution (CDC) and atrous spatial pyramid pooling (ASPP). CDC combines the pixel intensity information and the pixel gradient information to give a stationary description of texture difference information. Simultaneously, based on the ASPP, multi-scale information fusion can keep the texture features from being destroyed. Experimental results on several databases, Faceforensics++, DeeperForensics-1.0, Celeb-DF and DFDC prove that our MTD-Net outperforms existing approaches. The MTD-Net is more robust to image distortion, *e.g.*, JPEG compression and blur, which is urgently needed in the wild world.**

***Index Terms*—Face forgery detection, multi-scale, texture difference.**

I. Introduction

The rapid development of face manipulation techniques [1]–[3] has fueled the sharp increase of forgery face images and videos. These techniques, *e.g.*, DeepFakes [2], Face2Face [3], especially learning-driven generative models, such as Generative Adversarial Nets (GAN) [4], can create lifelike forgery face images and videos that cannot be distinguished even by human eyes. Some vivid examples are shown in Fig. 1. However, it is perilous that these methods are used for malicious purposes, *e.g.*, fake news, reputation infringement, even political purposes. Thus, it is extremely crucial to develop more effective approaches to face forgery detection.

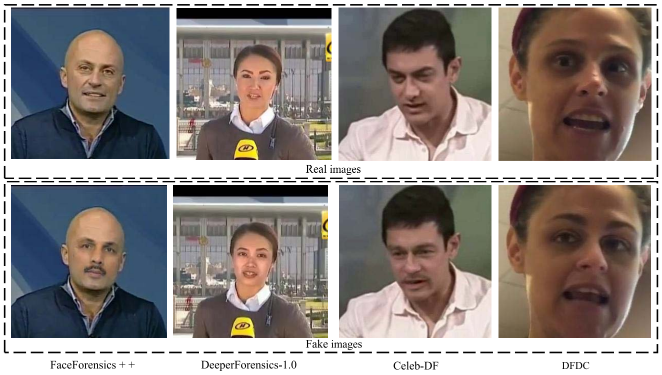


Fig. 1. Some vivid examples of real and fake images in the four databases, Faceforensics++, DeeperForensics-1.0, Celeb-DF and DFDC. The manipulation methods used in these databases are all based on deep learning. The real images are in the top row, and the fake images are in the bottom row.

In previous research, various methods [5]–[7] have been proposed for the most common traditional manipulations, such as splicing, copy-move, and removal. Most works utilize hand-crafted features, *e.g.*, illumination color [5], color filter array patterns [6], and blur type inconsistency [7], to classify a specific patch in an image as tampered or not. These features magnify the difference between real and fake images but focus on a single tampering technique. These methods have laid a solid theoretical foundation for image forensics research and inspired face forgery detection based on deep learning. However, in the face of today's widespread fraud, the direct use of these methods may be ineffective.

Deep learning has been broadly used in computer vision fields, such as object detection [8] in recent years. Scholars have also proposed some methods based on deep learning [9]–[12] to detect face forgery. Considering the influencing factors (distortion and compression) in the real environment, an explainable and robust model must adapt to this challenge.

Our work is inspired by two motivations. On the one hand, forgery clues may appear in different areas on the face. On the other hand, we want to know the behavior of CNN in identifying fake/real face images. Regarding the first motivation, we show some examples in Fig. 2, which indicate that forgery clues may appear in different areas on the face. Forgery clues are visually salient in occasional cases, for example, the lack of facial depth, the asymmetry of the

Manuscript received May 18, 2021; revised July 15, 2021; accepted July 16, 2021. Date of publication August 3, 2021; date of current version August 30, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61871283, in part by the Foundation of Pre-Research on Equipment of China under Grant 61400010304, and in part by the Major Civil-Military Integration Project in Tianjin, China under Grant 18ZXJMTG00170. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alessandro Piva. *(Corresponding author: Shuai Xiao.)*

Jiachen Yang, Aiyun Li, and Shuai Xiao are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: yangjiachen@tju.edu.cn; liaiyun@tju.edu.cn; xs611@tju.edu.cn).

Wen Lu is with the School of Electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: luwen@mail.xidian.edu.cn).

Xinbo Gao is with the Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: gaoxb@cqupt.edu.cn).

Digital Object Identifier 10.1109/TIFS.2021.3102487

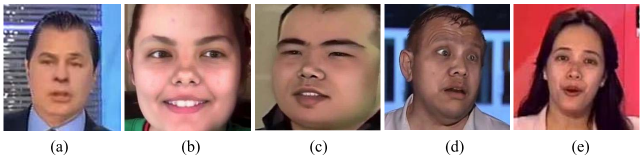


Fig. 2. Some examples of face forgery clues. (a) The lack of facial depth. (b) Asymmetry of the eyebrows. (c) The boundaries of a crop area can be seen on the right of the image. The examples in (d) and (e) are vivid such that no obvious face forgery clues can be seen.

eyebrows and the boundaries of a crop area shown in Fig. 2(a), Fig. 2(b), and Fig. 2(c). Human beings can quickly spot and utilize these clues to determine examples as "fake" even without further careful view. Nevertheless, human beings may give predictions with less confidence in most cases, as the clues mentioned above may be unnoticeable. For instance, no salient clue appears in Fig. 2(d), Fig. 2(e), and Fig. 1. Moreover, the fake faces look more vivid, and thus human beings cannot distinguish the difference with only a glance. However, CNN models can give an accurate prediction under these circumstances. Thus, to explore the behavior of CNN models may provide us useful information in facial forgery detection, which exposes our second motivation.

Under these two motivations, this study starts with the learning focus on CNN's behavior in the face forgery detection task, and we find that multi-scale texture difference information plays an important role in face forgery detection. Thus, a Multi-scale Texture Difference model (MTD-Net) is proposed to face forgery detection. Firstly, we use MTCNN [13] to crop faces from video frames and use a crop (extended by a factor of 1:3) around the center of the tracked face. Then, the crop faces are put into a module to extract the texture difference features. The pixel intensity information and the pixel gradient information are used to represent texture difference information. A special convolution operation is used to combine the intensity and gradient information. Next, we extract multi-scale information through a module. Finally, we fuse the extracted texture difference features at different scales for classification. The major contributions of our work can be outlined as follows:

- By studying the behavior of CNN, we realize that CNN has different regions of interest for real and fake images on face forgery detection, and the texture difference statistics of fake faces are different from real faces on multiple scales.
- We use the pixel intensity information and the pixel gradient information to give a stationary description of texture difference information. To extract the practical features, we leverage the advantage of a special convolution operation based on central difference convolution (CDC) [14]. To the best of our awareness, in the domain of face forgery detection, this is the first attempt to introduce special convolution operations for feature extraction and information fusion.
- In the process of combining multi-scale information, the use of atrous spatial pyramid pooling (ASPP) [15] and image-level pooling ensures that the information of the original features is not destroyed. Moreover, the loss of details due to the influence of dilated convolution is avoided.
- We propose a Multi-scale Texture Difference model (MTD-Net) for face forgery detection. Our model aims to extract and fuse the multi-scale texture difference information. Extensive experiments are conducted on three benchmark databases, Faceforensics++ [10], DeeperForensics-1.0 [16], Celeb-DF [17], and DFDC [18] to evaluate our method. Experimental results prove that our method performs better than other methods. Moreover, our method performs more robust in realistic data with high compression and mixed distortion. The proposed method effectively enhances the robustness and feasibility of a face forgery detection system.

The rest of this paper is arranged as follows: related works are briefly summarized in Section II. Empirical studies and analysis of CNN's behavior in the face forgery detection task are illustrated in Section III. The proposed method details are described in Section IV, and experiments and conclusions are presented in Section V and VI, respectively.

II. Related Work

A. Facial Manipulation Techniques

There are currently many public databases of real faces, such as CelebA, CASIA-WebFace, etc. The face data in these databases are obtained from public data used for research and does not invade personal privacy. As the facial manipulation techniques based on deep learning methods become more mature, researchers have established some public facial manipulation databases. For example, Faceforensics++, DeeperForensics-1.0, Celeb-DF, DFDC, etc. The quality of these facial manipulation databases is different, and the facial manipulation techniques used are also different in detail. However, they are all based on the data generated by the following facial manipulation techniques.

1) Deepfakes: The term *Deepfakes* has become the name of a specific facial manipulation technique. This technique aims to replace the faces of the target sequence with the faces of source videos or images. Currently, there are many open-source implementations of this method, such as in GitHub [2] and Fakeapp. This method uses a shared encoder and two automatic decoders. The shared encoder is trained to encode the features of the source face and the target face, and the automatic decoder is trained to reconstruct the training images of the source face and the target face. After that, the decoders corresponding to the source face and the target face are exchanged, and the resulting model can swap the source face and the target face.

2) NeuralTextures: NeuralTextures [19] is a facial manipulation technology based on the neural textures rendering method to display face reproduction. It learns the neural texture of the target face from the original video data and trains the rendering network corresponding to the target face. During the training process, the network uses a photometric reconstruction loss in combination with an adversarial loss. And the database Faceforensics++ uses a patch-based GAN-loss as used in Pix2Pix [20] to achieve a perfect reconstruction effect.

3) FaceSwap: FaceSwap [1] is a graphics-based method that can transfer facial regions from the source image to the target image. It extracts facial areas based on sparsely detected facial landmarks. The process uses these landmarks and blends shapes to fit the 3D template model. To get better results, researchers have proposed some methods combined with GAN. Researchers have added adversarial loss and perceptual loss to Variational Auto-Encoder (VAE) [21], which effectively improved the generation effect and solved the image blur and video jitter. Wang *et al.* [22] used CycleGAN [23] to improve the Pix2Pix method, which significantly enhanced the clarity and details of generated face images.

4) Face2Face: Face2Face [3] is a face reproduction system that can reproduce the facial expressions of the characters in the source video onto the faces of the characters in the target video while maintaining the identity information of the target character. In other words, Face2Face can exchange only the facial expressions of the characters. This method is based on two video input streams, manually select the keyframes in the video stream, extract the face information of the character, reproduce the face through calculation, transfer the expression through the mapping relationship, and re-synthesize under different conditions (lighting and expression) human face.

B. Forensics Methods

Inspired by the field of traditional image forensics and steganalysis, scholars have proposed some methods based on standard statistical features to detect facial tampering. In recent years, detection methods based on deep learning have shown their advantages in face forgery detection.

In traditional image forensics methods, most methods are based on statistical data or based on hand-designed functions. For the conventional image tampering mentioned in [24], scholars have proposed a series of effective detection methods. Lukas *et al.* [25] used the uniqueness of the correlation between the camera's fixed pattern noise and the source device to detect tampered images. Cozzolino *et al.* [26] proposed to detect image splicing through summarized noise statistics. Stamm and Bayar [27] proposed constrained convolutional neural networks to suppress image content. This method provided an important foundation for subsequent forensic research. Light sources were used to calculate the direction of scene lighting in [28] and the inconsistencies in lighting were used to determine whether the image has tampered. Other methods analyzed, for example, JPEG compression artifacts [29], [30] and traces of resampling [31]. Fridrich and Kodovsky [32] used a hand-made function to scan features with a pixel radius of 2 along with the horizontal and vertical directions of the image using a high-pass filter, and used these features to train a linear Support Vector Machine (SVM) classifier. Some detection methods based on neural networks were proposed, for example, setting noise patterns on EXIF entries [33] or directly searching for unqualified noise [34]. Zhou *et al.* [35] proposed a dual-stream network which can detect tampered images and distinguish the tampering methods. These methods have laid a solid theoretical foundation for image forensics research and inspired forged image detection based on deep learning. However, with the improvement of GANs, the application of traditional methods in deep forgery detection becomes more and more difficult.

Since the forgery faces become more and more realistic, some works [36], [37] utilized deep networks to learn discriminative features or find manipulation traces to detect face forgery. Li *et al.* [38] found that human eyes of GAN-based videos did not blink. A simple deepfakes detection network MesoNet was created in [39], which gave some interpretable conclusions. Rossler *et al.* [10] introduced an effective Xception-Net as a binary face forgery image detector. Li *et al.* [40] focused on artifacts or splicing traces produced to generate tampered images and achieved great results. Amerini *et al.* [41] used the difference of optical the flow field as a clue to identify deepfakes video and original video, considering possible anomalies in the time dimension of the sequence. Durall *et al.* [42] transferred the research field from the spatial domain to the frequency domain, used the power spectrum as a feature for forensic image forensics. Two features based on frequency domain design were proposed in [43], called Frequency-aware Decomposition (FAD) and Local Frequency Statistics (LFS). Tolosana *et al.* [44] provided an exhaustive analysis of both 1st and 2nd DeepFakes generations in terms of facial regions and fake detection performance, which provided a good reference for follow-up research. Liu *et al.* [45] focused on global texture and introduced the gram module into the network. This method significantly improved the robustness and provided a direction for subsequent research. Kumar *et al.* [46] proposed a approach based on metric learning, which provided an important foundation for subsequent classifier research in face forgery detection. Some recent works used biological information [47], such as heartbeat information [48] to detect deepfakes videos.

III. Empirical Studies and Analysis

This section starts with the behavior of CNN models in distinguishing facial forgery, and we realize that CNN has different interest areas for real and fake images through visualization experiments. Based on the above observation, some experiments are designed for further analysis.

A. Visual Perception

The CNN models can already achieve excellent results in forgery detection tasks. However, the data-driven CNN models lack interpretability. Moreover, the robustness of the data-driven CNN models is generally very poor, and the performance on the distorted data such as (distortion and compression) is also not good enough. Thus, to understand the behavior of CNN in forgery detection tasks deeply, we first design a visualization experiment to understand the interest areas of the CNN models in the forgery detection task.

The contribution map CM^c is calculated to reflect which parts of the input the CNN model's attention is on. A simple CNN model, ResNet-18 [50], was trained on real and fake face images from Faceforensics++ (DFc23). Assuming there is a face image I, we put it into our trained ResNet-18 model. Then, we get the prediction c (the value of c is 0 or 1, indicating the fake or real face image) and corresponding

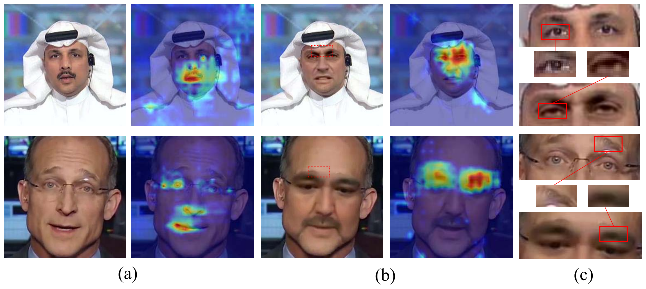


Fig. 3. The visual images obtained using the method in [49]. The red regions represent excellent attention, and the blue regions represent insufficient attention. The left side is the input image, and the right side is the Grad-CAM image. (a) Real face images. (b) Fake face images. (c) Areas comparison between the real images and fake images.

output value before the softmax layer y^c. We take out K feature maps from the last convolutional layer. The pixel value at position (i, j) in the k^{th} feature map is defined as $I^k_{(i,j)}$. The weights of the k^{th} feature map to the class c can be obtained using Gradient-weighted Class Activation Mapping (Grad-CAM) [49]:

$$w_k^c = \sum_i \sum_j \frac{\partial y^c}{\partial I^k_{(i,j)}} \tag{1}$$

The contribution map CM^c is as follows,

$$CM^c = \frac{1}{K} \sum_k w_k^c I^k \tag{2}$$

the contribution maps are shown in Fig. 3.

We find that the CNN model has different focus areas for real and fake face images in face forgery detection. For real face images, the CNN model pays more attention to the mouth and nose regions. For fake face images, the CNN model focuses more on the eyes and eyebrows regions. By visually observing the focus areas of fake image, the intuitive difference between the real and fake face images can be easily found.

As shown in Fig. 3(c), the texture of the eyes regions in the real face image is obvious, while the surface of the eyes regions in the fake face image is very blurry. This situation also occurs in eyebrows areas with great attention. We expand the observation scope and find that this “fuzzy” phenomenon only appears in the facial area; in other words, the tampered area. We can even observe the border caused by texture differences, marked in Fig. 3(b). The above observations prompt us to study further whether the fake face is different from the real face in terms of texture difference statistics.

B. Regional Differences

Texture is an image feature formed by the repeated occurrence of gray pixel levels in space. To study the texture differences between real face images and fake face images, we use Gray Level Co-occurrence Matrix (GLCM) [51] to perform a qualitative analysis.

The GLCM $P_\theta^d \in R^{256\times256}$ represents the co-occurrence of measured pixel values under a given offset parameterized by distance d and angle θ. $P_\theta^d(i, j)$ means how often a pixel with value i and a pixel at offset (d, θ) with pixel value j co-exist. In the qualitative analysis, we calculate P_θ^d separately on face images and then calculate the statistics that can intuitively represent texture information based on P_θ^d. We select parameter $d \in \{1, 2, 5, 10, 15, 20\}$, parameter $\theta \in \{0, \pi/2, \pi, 3\pi/2\}$. The texture contrast is chosen as a statistic to measure texture information more intuitively. The texture contrast C is calculated under different distance offsets. Various d and θ combinations ensure that the complete pixel relationship information of the face image. The formula for texture contrast is expressed as follows,

$$C = \frac{1}{N} \sum_{i,j=0}^{255} \sum_{\theta=0}^{3\pi/2} |i - j|^2 P_d^\theta(i, j) \tag{3}$$

where N is a normalization factor and $N = 256 \times 256 \times 4$. i, j represents the intensity of the pixels. A low C value indicates blurry texture. On the contrary, larger C reflects stronger texture contrast and clearer visual effects. The contrast component C is shown in TABLE I.

TABLE I

THE VALUE OF C WITH DIFFERENT DISTANCE d. REAL FACES RETAIN LARGER TEXTURE CONTRAST THAN FAKE FACES AT ALL MEASURED DISTANCES

Database \ Distance	1	2	5	10	15	20
Realc23	1.90	4.15	12.08	20.55	27.73	35.06
DFc23	1.37	3.59	9.65	17.91	25.04	31.36
FSc23	1.43	3.75	9.75	17.99	25.82	31.47

We use 1,000 face images from the Faceforensics++ database (c23), randomly selected for each category, to calculate the parameter C. It can be seen that at all distances d, real faces retain larger contrast than fake faces. We notice that CNN-based generators usually normalize the values of associated pixels during the generation process, which results in images generated based on CNN that cannot restore the texture contrast as strong as the real images. This difference is reflected in multiple sizes of d, and as d increases, the difference in texture becomes more apparent. In this paper, we only conduct a quantitative analysis of the texture contrast. The results show that multi-scale texture difference information can play a role in face forgery detection.

IV. PROPOSED METHOD

In this work, we propose a multi-scale texture difference model inspired by how CNN models can behave to anticipate whether a face image is real or fake. In Section IV-A, we describe how the CDC work to extract texture difference features. Subsequently, in Section IV-B, we illustrate the extraction of multi-scale information from texture difference feature map F_T. Finally, in Section IV-C, the whole network architecture and the training process are presented.

A. Texture Difference Features Extraction

The texture difference features extraction aims to exploit textural discriminative information. In previous work,

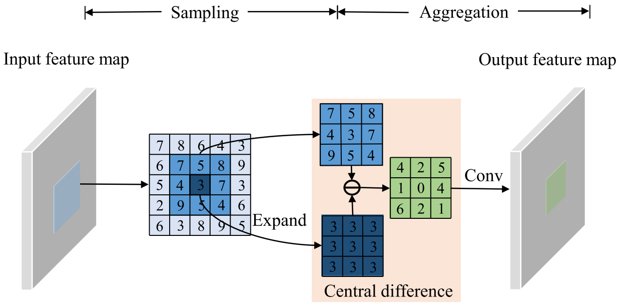


Fig. 4. The process of CDC.

Gram-Net [45] used the gram matrix to extract global texture difference information of the full image generated by GAN. However, a new description of texture difference information is needed for the face of locally generated images. In our work, following the idea of CDCN [14], we use the pixel intensity information and the pixel gradient information to give a stationary description of texture difference information.

The most basic vanilla convolution is used to extract the pixel intensity information. The process of vanilla convolution extracting the next-level feature map F_{l+1} from the feature map F_l is as follows,

$$F_{l+1}\left(p_c\right)=\sum_{p_r \in r} w\left(p_r\right) F_l\left(p_c+p_r\right) \tag{4}$$

where p_c is a current location on both input feature map F_l and feature map F_{l+1}. r is the local receptive field region for vanilla convolution operation, and p_r enumerates the locations in r. For instance, $p_r \in \{(-1,-1),(-1,0),\cdots,(0,1),(1,1)\}$ with the 3×3 kernel and dilation 1, and $w(p_r)$ stands for the weight.

For the pixel gradient information, we use CDC to give a stationary description. The CDC combines the idea of difference with the convolution operation to enhance the ability of features to express pixel gradient information. The process of CDC extracting the next-level feature map F_{l+1} from the feature map F_l is as follows.

$$F_{l+1}\left(p_c\right)=\sum_{p_r \in r} w\left(p_r\right)\left(F_l\left(p_c+p_r\right)-F_l\left(p_c\right)\right) \tag{5}$$

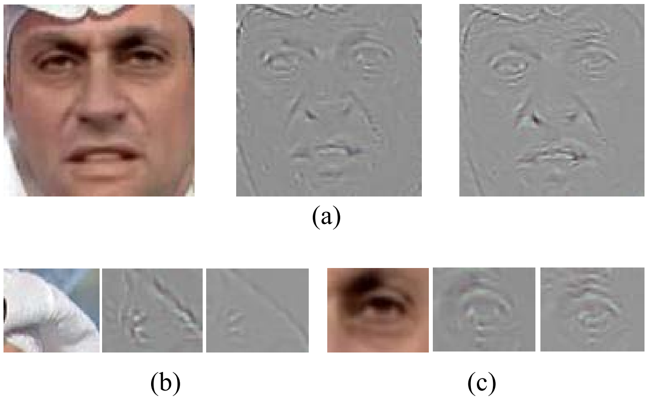
The details are shown in Fig. 4. Then, we combine vanilla convolution and the CDC with a parameter $\alpha \in [0,1]$.

$$\begin{aligned} F_{l+1}\left(p_c\right)=&(1-\alpha) \sum_{p_r \in r} w\left(p_r\right) F_l\left(p_c+p_r\right) \\ &+\alpha \sum_{p_r \in r} w\left(p_r\right)\left(F_l\left(p_c+p_r\right)-F_l\left(p_c\right)\right) \end{aligned} \tag{6}$$

The formula after decomposition and merger is as follows.

$$F_{l+1}\left(p_c\right)=\sum_{p_r \in r} w\left(p_r\right) F_l\left(p_c+p_r\right)-\alpha \sum_{p_r \in r} w\left(p_r\right) F_l\left(p_c\right) \tag{7}$$

We call the generalized central difference convolution described by Eq. 7 as the full version of CDC, and refer to the work [14] in face anti-spoofing (FAS), we set $\alpha = 0.7$. It is worth noting that CDC does not increase the number of network parameters in the specific implementation process. We replaced the convolutions of ResNet-18 by central difference convolutions, and trained the model under the same setting mentioned in Section III-A. We extracted the features of the last convolutional layer in the convergent model. As shown in Fig. 5(a), (b), and (c), from left to right are the fake image, the features extracted by CDC, and the features extracted by VC. It can be seen from Fig. 5(b), the edges of the character's scarf represent the real area in the image, the features of CDC are stronger than the features of VC. And the eyes of the character represent the fake area in the Fig. 5(c), the features of CDC reflect the actual information of the image more than the features of VC. The feature map extracted using CDC can better reflect the essential features of forged face images.


(a)

(b)

(c)

Fig. 5. The visual image obtained using the method in [52]. We can clearly see that the extracted features are different. From left to right are the fake image, the features extracted by CDC, and the features extracted by VC. (a) Full face. (b) Edge of scarf (real area). (c) Eyes (fake area).

B. Multi-Scale Information Extraction

Multi-scale information is expected to exploit discriminative information from different-scale features. We use a special convolution operation, dilated convolution [53], to extract different-scale features from the feature map F_T, which is the output of texture difference features module. The process of dilated convolution is as follows,

$$M_d\left(p\right)=\sum_{t \in r} F_T(p+dt) w(t) \tag{8}$$

where $M_d\left(p\right)$ is the output of dilated convolution, and p is a current location on M. $w(t)$ is a 3×3 kernel, t is a location in the kernel. p is the location in the feature map F_T, and d is the dilation rate. Inspired by the previous work [15], we use dilated convolution with different dilation rates $d \in \{6, 12, 18\}$ to extract different-scale information.

To avoid the loss of local texture information, we refer to the work [15] introduces ASPP and image pooling to better integrate different-scale information. The process of image pooling is as follows,

$$M_{\text{Image}}=U\left(\delta\left(F_T\right)\right) \tag{9}$$

where δ is the 1×1 convolution, and U is the upsampling operation. The output of the final multi-scale information

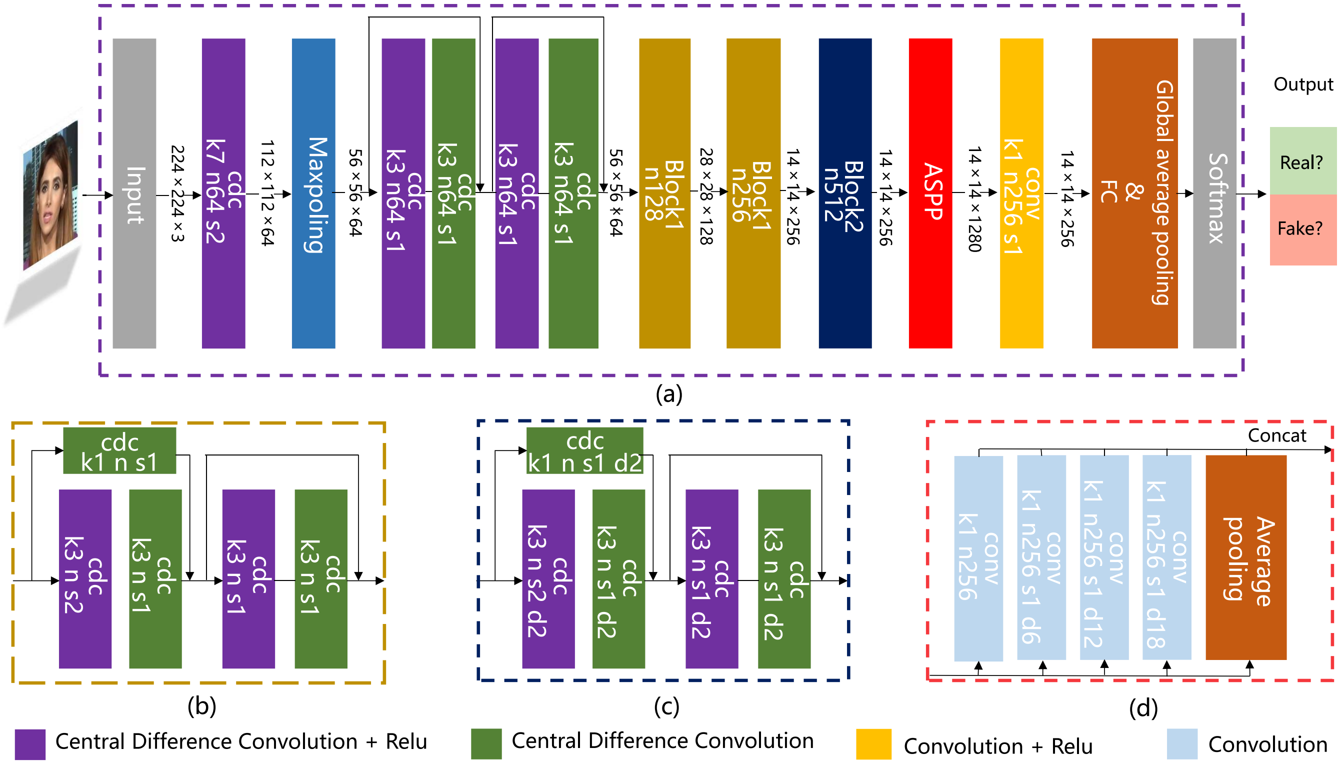


Fig. 6. The architecture of our MTD-Net: (a) the architecture of our network, comprised of a texture difference feature module (marked with the purple dotted line) and a multi-scale information module (marked with the green dotted line), (b) the block1 in texture difference feature module, (c) the block2 in the in texture difference feature module, and (d) the ASPP block in the multi-scale information module. In particular, the "k" represents kernel size, the "n" represents the number of channels, the "s" represents stride size, and the "d" represents the dilation rate.

fusion is as follows.

$$M = \delta(\text{concat}(M_{\text{Image}}, [\delta(F_T), M_6, M_{12}, M_{18}])) \quad (10)$$

C. The Networks for Face Manipulation Detection

Our work aims to train a function $Y_{\text{fake/real}} = P(x_{\text{face}})$, which gives an accurate prediction whether the input face image real or fake. So we propose a multi-scale texture difference model MTD-Net.

The proposed network is comprised of the texture difference feature module and multi-scale information module. As shown in Fig. 6, the texture difference feature module is built based on the a simple 18-layer ResNet, since the shortcuts can combine intuitive features near the bottom layer and abstract features near the top layer. The convolutions of the basic ResNet-18 are replaced by CDC to make full use of the texture difference features. In addition, to extract enough features for face forgery, we adjust the size of the feature map by setting dilation rate to 2 in block2. The extracted features are integrated in the next module.

After the texture difference feature module, we propose a multi-scale information module, which consists of ASPP block, to extract multi-scale information. The structure of the module is inspired by the work in [15]. The ASPP block performs parallel dilated convolution with different dilation rates to the input feature map in order to capture different scale information, and then fuses them together. Then, a 1 × 1 convolution layer is used to learn the adaptive re-calibration of the extracted features. After the multi-scale information module, we use utilize global average pooling to squeeze the spatial information into channel statistics and sent the feature information to the fully connected layer for final classification. The detail settings for the full MTD-Net are shown in Fig. 6 (a), (b), (c), and (d). The visual images are shown in Fig. 7 (a), (b), and (c). We can clearly see that compared with the basic network ResNet-18, the proposed MTD-Net uses a larger range features to make judgments, which includes the real area and the fake area. The experimental results also show that MTD-Net have a good effect on face forgery detection.

Considering the difference between the two tasks, we use random initialization for the lays and blocks in face forgery detection, which is different from the specific initialization used for the dilated convolution in the image segmentation [53]. The networks are optimized via Adam [54]. We set the base learning rate as 0.001 and use Cosine [55] learning rate scheduler, and the momentum is set as 0.9. The batch size is set as 64 for about 50 epochs training. For the loss function, we choose the cross-entropy function, which is often used as the loss of binary classification task functions,

$$L(y_t, y_p) = -(y_t \times \log(y_p) + (1 - y_t)\log(1 - y_p)) \quad (11)$$

we assume that the true label of y is y_t, and the probability of $y_t = 1$ is y_p, where label $\in \{0, 1\}$.

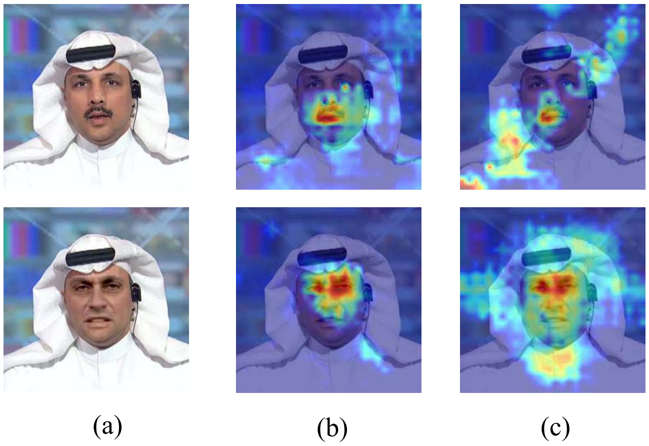


Fig. 7. The visual images obtained using the method in [49]. The red regions represent excellent attention, and the blue regions represent insufficient attention. (a) The input face images. (b) ResNet-18. (c) MTD-Net.

V. Experiments and Results

In this section, extensive experiments are conducted to verify the effectiveness of our proposed method. We first briefly introduce three benchmark databases. Then, implementation details are illustrated. Finally, we present and discuss our experimental results.

A. Databases

In this subsection, to assess the effectiveness of our method, an experimental evaluation on Faceforensics++, DeeperForensics-1.0, Celeb-DF and DFDC is provided. Fig. 1 shows some examples.

1) Faceforensics++: Faceforensics++ [10] is a large-scale facial manipulation database that consists of real portrait videos and fake portrait videos. Most real portrait videos are collected from YouTube with the consent of the subjects. Each real portrait video undergoes four manipulation methods, *i.e.*, Deepfakes, FaceSwap, Face2Face, and NeuralTexture, to generate four fake videos. Each manipulation method contains 1,000 videos. Output videos are developed with three quality levels, *i.e.*, raw, c23, and c40, corresponding to high quality, medium quality, and low quality, respectively. We followed the set of previous work [10] to partition the database to compare with other methods. For 1,000 videos in each sub-database, we used 720 videos for training, 140 videos for validation, and 140 videos for testing. We sampled 270 frames from each training video, and 100 frames from each validation and testing video. Our performance report was achieved on c23 and c40.

2) DeeperForensics-1.0: DeeperForensics-1.0 [16] is a large-scale database for real-world face forgery detection. The database collects face data from 100 individuals and takes 1,000 refined YouTube videos collected by Faceforensics++ as target videos. Each face of the ordered 100 identities is swapped onto ten target videos by an end-to-end process. Besides, videos add various perturbations, which are mentioned in Image Quality Assessment (IQA) [56], [57], to simulate videos in real scenes. We followed the set in previous work [16] to partition the database. For 1,000 videos in each sub-database, videos were split into training, validation, and testing with 7: 1: 2. We sampled 270 frames from each training video and 100 frames from each validation and testing video. Our performance report was achieved on std, std/random, and std/mix.

3) Celeb-DF: Celeb-DF [17] is a new large-scale and challenging deepfakes video database. The Celeb-DF database aims to generate fake videos of better visual quality compared with the previous database. This database contains 590 real videos extracted from YouTube, matching celebrities of various gender, age, and ethnic groups. These videos exhibit an extensive range of aspects, such as face sizes, lighting conditions, and backgrounds. As for fake videos, a total of 5,639 videos are created swapping faces using deepfakes technology. The final videos are in MPEG4.0 format. We followed the set in previous work [59] to partition the database and sampled 32 frames for each video. We use the test set provided by the database itself, and we randomly select 15% of the videos as validation set, with the remaining 85% for training.

4) DFDC: DFDC [18] is a large publicly-available face swap video database, with over 100,000 total clips sourced from 3,426 paid actors, produced with several Deepfake, GAN-based, and non-learned methods. We followed the set in previous work [60] to partition the database and sampled 32 frames for each video. We split DFDC according to its folder structure, using the first 35 folders for training, folders from 36 to 40 for validation and the last 10 folders for testing.

B. Experimental Setups

In this subsection, the experimental settings of our method is presented so that the other researchers can reproduce our results.

1) Setting: The hardware conditions of the experiments are Intel (R) Xeon (R) CPU E5-2620 V4 and two NVIDIA GTX Titan XP GPUs. The input crop face images are resized to a fixed size of 224×224 pixels before inputting into the network. The model updates and saves through the hyperparameters setting.

2) Evaluation Metrics: We report the Accuracy score (ACC) and Area Under the Receiver Operating Characteristic Curve (AUC) as our evaluation metrics.

a) ACC: We use the frame-level ACC as a significant evaluation of Faceforensics++ and DeeperForensics-1.0. The ACC formula is as follows,

$$ACC = S_{tp}/ALL_{test} \tag{12}$$

where S_{tp} is the number of images classified correctly, ALL_{test} is the number of pictures participating in the test.

b) AUC: AUC is used as another evaluation metric for Celeb-DF and DFDC. We computed binarizing the network output with different thresholds to calculate the frame-level AUC.

C. Performance Evaluation

In this section, the performance of the proposed method is analyzed and compared with other state-of-the-art ones.

TABLE II

COMPARATIVE ANALYSIS OF DETECTION PERFORMANCE WITH THE OTHER RECENT METHODS. EACH DETECTION METHOD IN THE TABLE IS TRAINED AND TESTED UNDER THE CONDITION OF FACE IMAGES

Database	Faceforensics++ c23 [ACC]				Faceforensics++ c40 [ACC]				Deeperforensics 1.0 [ACC]		Celeb-DF [AUC]	DFDC [AUC]
Method	DF	F2F	FS	NT	DF	F2F	FS	NT	std/random	std		
Durall *et al.* [42]	81.72	89.26	90.33	72.73	71.69	65.66	65.43	59.34	79.53	85.61	0.7846	-
Rahmouni *et al.* [36]	82.16	93.48	92.51	75.18	73.25	62.33	67.08	62.59	80.38	87.55	0.8127	-
MesoNet [39]	95.26	95.84	93.43	85.96	89.52	84.44	83.56	75.74	95.76	94.77	0.9863	0.7887
XceptionNet [10]	98.85	98.36	98.23	94.50	94.28	91.56	93.7	82.11	99.63	**100**	0.9947	0.8589
DSP-FWA [58]	96.57	97.33	95.83	91.51	93.60	91.77	90.73	83.15	99.09	98.24	0.9913	0.8526
Liu *et al.* [45]	97.63	96.31	97.96	92.07	92.39	90.67	91.99	84.69	99.25	99.71	0.9866	0.8477
Qian *et al.* [43]	98.43	98.51	98.34	93.22	96.01	93.62	94.33	86.37	98.17	98.89	0.9969	0.8843
Bondi *et al.* [59]	97.55	98.73	98.29	93.57	94.95	91.33	94.26	87.79	99.37	99.89	0.9980	0.9190
Bonettini *et al.* [60]	**98.94**	99.34	98.04	94.53	96.13	92.93	94.09	88.15	99.72	99.98	**0.9991**	**0.9226**
Proposed	98.64	**99.76**	**98.42**	**94.60**	**97.88**	**96.85**	**96.86**	**88.47**	**99.91**	99.93	0.9987	0.9197

We test the performance of the proposed method on three databases. The results are shown in TABLE II.

In the comparison part, we chose some classic methods, which are based on frame-level detection. Durall *et al.* [42] proposed a method using the high-frequency information difference to train SVM for identification. Rahmouni *et al.* [36] adopted different CNN architectures with a global pooling layer that computes four statistics (mean, variance, maximum, and minimum). MesoNet [39] is a CNN-based network to detect face forgery. XceptionNet [10] is a traditional CNN trained on ImageNet based on separable convolutions with residual connections. DSP-FWA [58] employed a dual spatial pyramid strategy to tackle multi-scale issues, and the use of multi-scale information provides important inspiration for follow-up research. Liu *et al.* [45] focused on global texture and introduced the gram module into the network. Qian *et al.* [43] proposed a method based on frequency domain. Bondi *et al.* [59] used EfficientNet B4 for more accurate detection, which provided a good training strategy in face forgery detection. Bonettini *et al.* [60] proposed a model based on EfficientNet, which used attention layers. In this paper, we do not use any unique training method for the fair comparison, *i.e.*, special loss. The input of these methods is only a single face image without any additional input, *i.e.*, masks or other information. For all methods whose source codes are opened to the public, we conducted experiments for them by ourselves.

On the Faceforensics++ database, the proposed MTD-Net achieved great results. The ACC in some categories exceeds the reference methods in c23, *i.e.*, F2F, FS, NT. The ACC in all categories exceeds the reference methods in c40. A possible explanation for this result might be that the MTD-Net performs better on low-quality data. Moreover, the ACC of the proposed method decreases less than other reference methods when the compression level increases from c23 to c40.

On the other three databases, the proposed method can generally achieve great performance. On DeeperForensics-1.0 database, our MTD-Net gets the best results in std/random, and this finding suggests that the proposed method shows stronger robustness on distorted data. Our MTD-Net ranks second among all methods on Celeb-DF and DFDC, indicating that our method has good generalization performance in face forgery detection. From the TABLE II, one can see that the proposed MTD-Net obtains the best overall performance.

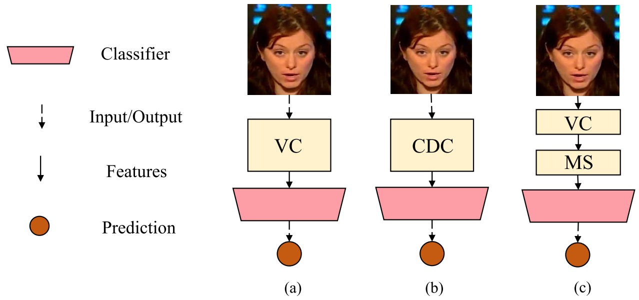


Fig. 8. Algorithmic configuration in module comparison. VC represents the feature extraction layer in ResNet-18 using vanilla convolution. CDC defines the feature extraction layer in ResNet-18 using CDC. MS represents multi-scale extraction and fusion. (a) Resnet. (b) Res_CDC. (c) Res_MS.

D. Module Comparison Evaluation

This section validates the contribution of multi-scale information and texture difference features to the proposed method. For comparison purposes, we use the most basic network as ResNet-18. ResNet-18 does not include any multi-scale information module or texture difference features module. For convenience, we refer to this model as “Resnet” in the following. On this basis, the version with multi-scale information is represented as Res_MS, and the version with texture difference features as Res_CDC. An intuitive algorithm configuration is shown in Fig. 8.

During the module comparison experiment, we first designed independent experiments to discuss the role of multi-scale information and texture difference features, respectively. Then, we conducted module reduction experiments on three databases. The experiment configuration under the same investigation remains unchanged. Only the modules that need to be compared are added or deleted.

TABLE III

COMPARATIVE ANALYSIS OF THE BENEFITS OF MULTI-SCALE INFORMATION. EACH DETECTION METHOD IN THE TABLE IS TRAINED AND TESTED UNDER THE CONDITION OF FULL IMAGES

Database	Faceforensics++ c23				
Method	DF	F2F	FS	NT	Real
Xception	88.00	84.98	82.23	79.60	65.85
Resnet	89.78	81.54	78.90	63.06	82.37
Res_MS	**97.70**	**92.80**	**89.80**	**86.09**	**93.98**

TABLE IV

COMPARATIVE ANALYSIS OF THE BENEFITS OF TEXTURE DIFFERENCE FEATURES. EACH DETECTION METHOD IN THE TABLE IS TRAINED AND TESTED UNDER THE CONDITION OF FACE IMAGES

Database	Faceforensics++ c23				
Method	DF	F2F	FS	NT	Real
Resnet	92.91	97.01	95.93	60.99	93.48
Res_CDC	**98.61**	**98.84**	**97.80**	**93.39**	**97.99**

TABLE V

COMPARATIVE ANALYSIS OF THE BENEFITS OF TEXTURE DIFFERENCE FEATURES. EACH DETECTION METHOD IN THE TABLE IS TRAINED AND TESTED UNDER THE CONDITION OF FACE IMAGES

Method Database		Resnet	Res_CDC
Train	Test		
std	std	98.17	**99.29**
std/random	std/random	98.06	**99.86**
std/mix	std/mix	97.56	**99.69**
std	std/random	92.08	**95.22**
std	std/mix	75.43	**86.89**
std/random	std	93.96	**97.81**
std/random	std/mix	99.73	**99.91**
std/mix	std	35.61	**62.66**
std/mix	std/random	70.63	**88.81**

TABLE VI

COMPARATIVE ANALYSIS OF MODULE COMPARISON EVALUATION. EACH DETECTION METHOD IN THE TABLE IS TRAINED AND TESTED UNDER THE CONDITION OF FACE IMAGES

Method Database	Resnet	Res_CDC	Res_MS	Proposed
Realc23	93.48	97.99	98.33	**98.49**
DFc23	92.91	98.61	98.45	**98.64**
F2Fc23	97.01	98.84	99.62	**99.76**
FSc23	95.93	97.80	98.38	**98.42**
NTc23	60.99	93.39	94.01	**94.60**
std	97.56	99.29	99.40	**99.81**
Celeb-DF	97.63	99.15	99.27	**99.71**

1) The Benefits of Multi-Scale Information: To further prove the benefits of multi-scale information in face forgery detection, we designed an experiment on full images in Faceforensics++ c23. The information of "full image" is more prosperous, and the "full image" scan can better reflect the advantages of multi-scale information. The setting is the same as that of performance evaluations. As shown in the last row of TABLE III, the ACC of Res_MS is much higher than Xception. Moreover, the ACC of each category has been dramatically improved compared with the primary network Resnet. For face images, we only compared with the primary network in this paper. The results are shown in TABLE VI. We can see the difference between column 2 and column 4 in TABLE VI, proving that multi-scale information is beneficial in face forgery detection.

2) The Benefits of Texture Difference Features: To verify the texture difference features in face forgery detection, we designed independent experiments on Faceforensics++ c23 and DeeperForensics-1.0. In this subsection, we aim to evaluate the robustness of the texture difference features. The experiments were conducted on face images, and the setting was the same as that of performance evaluations.

On the Faceforensics++ c23, we did verification tests on different categories. As shown in TABLE IV, texture difference features help improve ACC compared with the primary network Resnet. To further evaluate the feasibility of texture difference features, an experiment was designed on distorted data in DeeperForensics-1.0. The detailed configuration is shown in TABLE V. The hyperparameters setting is the same as that of the previous work [16]. We used manipulated data in the standard set (std), manipulated videos with random-type, random-level distortions (std/random), and manipulated videos with a mixture of three random-level, random-type distortions (std/mix).

In TABLE V, the ACC of the two methods is high when the models are trained and tested on the standard database. It is expected that the models perform well on high-quality databases. It is worth noting that the ACC of Resnet decreases when we choose distorted databases. However, the ACC of Res_CDC increases on the distorted databases, and it performs better than the Resnet. It seems possible that these results are due to the texture difference features have better representation, which is more robust.

To further evaluate the performance, we swapped the training data and the test data. The detailed configuration is shown in TABLE V. The ACC of Resnet is significantly affected, while the impact on Res_CDC is small. The experimental results reflect the robustness of the texture difference features. Moreover, in rows 3 and 4, the effects on both methods are small. Initial observations suggest that there may be a link between the quality of training data and accuracy.

3) The Benefits of Fusion Information: In this subsection, we prove the performance of the proposed method by jointly exploiting texture difference features and multi-scale information. Module comparison evaluations were conducted on Faceforensics++ c23, std (DeeperForensics-1.0) and Celeb-DF. The results are shown in TABLE VI, where the proposed represents the complete method MTD-Net.

As shown in TABLE VI, the results with only multi-scale information exceed the results with superior texture difference

TABLE VII

COMPARATIVE ANALYSIS OF CROSS-DATABASES EXPERIMENT. EACH DETECTION METHOD IN THE TABLE IS TRAINED AND TESTED UNDER THE CONDITION OF FACE IMAGES

Database	Method	MesoNet	XceptionNet	Durall *et al.*	Bondi *et al.*	Bonettini *et al.*	Proposed
Train	Test						
F++c23	F++c23	0.9876	0.9894	0.8544	0.9600	0.9911	**0.9938**
std/random	std/random	0.9962	0.9977	0.7890	0.9994	**0.9999**	**0.9999**
Celeb-DF	Celeb-DF	0.9763	0.9927	0.7346	0.9980	**0.9991**	0.9987
F++c23	std/random	0.5766	0.7003	0.5289	0.7349	0.7345	**0.8421**
F++c23	Celeb-DF	0.6488	0.6712	0.5434	**0.7340**	0.7114	0.7012
std/random	F++c23	0.6124	0.5163	0.4392	0.5849	0.5478	**0.6575**
std/random	Celeb-DF	0.5099	0.4691	0.4235	0.5261	0.5164	**0.5343**
Celeb-DF	F++c23	0.6124	0.5824	0.4739	0.6150	**0.6385**	0.6343
Celeb-DF	std/random	0.4636	0.4850	0.4572	0.5873	0.5567	**0.6071**

features. One possible implication is that multi-scale information may be more efficient than texture difference features. Intuitively, human are likely to give a reliable judgment with a glance at the apparent artifacts, such as the asymmetry of the eyebrows or the boundaries of crop area. The artifacts are located in different locations and have different scales, corresponding to multi-scale information. However, when the discrimination artifacts are not noticeable, the subtle texture difference features may play a role, which we found when discussing the behavior CNN model. Besides, by fusing texture difference and multi-scale information, the proposed method can further achieve better performance. This improvement supports our motivation that "multi-scale texture difference features" can provide more details to improve the classification.

E. Cross-Database Experiment

To further verify the performance of the proposed method, a cross-database experiment was designed. We trained and test on F++c23 (DF, F2F, FW, NT and Real), std/random (DeeperForensics-1.0) and Celeb-DF. We followed the set in previous work [59] to partition the database. We only consider the benefits of the network structure, and the input of the network is only a single face image, without any other information. To comparison with other methods easily, we use AUC as an evaluation indicator. Detailed configurations and results are shown in TABLE VII.

From the experimental results, we can see that if the training set and testing set are the same data distribution, the detection effects of all methods are excellent. However, the AUC is significantly reduced when the training set and the testing set belong to different data distributions.

We can still dig out useful information from the results. The model trained with F++c23 is more effective on the test set of other databases than the model trained on std/random and Celeb-DF. Experimental results show that training the model with data that has more fake category distributions might improve generalization ability. Moreover, training the model with a high degree of distortion might increase the robustness of the model. Our MTD-Net has a strong ability to deal with data distortion. MTD-Net can still reflect certain advantages in cross-database experiments.

The distribution difference between the databases is big. The face forgery databases must consider different methods of the forgery and evaluate the effects of environmental distortion. How to propose a highly generalized method in this data environment is also our future research direction.

VI. CONCLUSION AND PROSPECTION

Our study provided a method of face forgery detection based on multi-scale texture difference information. The texture differences between the real and fake face images were found by the traditional texture representation GLCM. Based on the finding, we first attempted to introduce particular convolution operations-CDC for feature extraction and information fusion. Meanwhile, the advantage of ASPP and image-level pooling was also leveraged to fuse multi-scale information. The evaluation experiments were conducted on the Faceforensics++, DeeperForensics-1.0, Celeb-DF and DFDC databases compared with the recent state-of-the-art model. Experiments prove that our method can achieve high accuracy and also perform well in combating distortion. However, it is still necessary to train a new model for an unknown face manipulation method. In future research work, we will commit to forming a universal approach and promoting face forgery detection.

REFERENCES

[1] Faceswap. (2019). *DeepFakes*. [Online]. Available: https://www.github.com/MarekKowalski/FaceSwap/

[2] DeepFakes GitHub. (2019). *DeepFakes*. [Online]. Available: https://github.com/deepfakes/faceswap

[3] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, "Face2Face: Real-time face capture and reenactment of RGB videos," *Commun. ACM*, vol. 62, no. 1, pp. 96–104, 2018.

[4] I. J. Goodfellow *et al.*, "Generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 3, 2014, pp. 2672–2680.

[5] T. J. de Carvalho, C. Riess, E. Angelopoulou, H. Pedrini, and A. Rocha, "Exposing digital image forgeries by illumination color classification," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 7, pp. 1182–1194, Jul. 2013.

[6] P. Ferrara, T. Bianchi, A. D. Rosa, and A. Piva, "Image forgery localization via fine-grained analysis of CFA artifacts," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1566–1577, Oct. 2012.

[7] K. Bahrami, A. C. Kot, L. Li, and H. Li, "Blurred image splicing localization by exposing blur type inconsistency," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 5, pp. 999–1009, May 2015.

[8] J. Wen, J. Yang, B. Jiang, H. Song, and H. Wang, "Big data driven marine environment information forecasting: A time series prediction network," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 1, pp. 4–18, Jan. 2021.

[9] D. Cozzolino, G. Poggi, and L. Verdoliva, "Recasting residual-based local descriptors as convolutional neural networks: An application to image forgery detection," in *Proc. 5th ACM Workshop Inf. Hiding Multimedia Secur.*, Jun. 2017, pp. 159–164.

[10] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.

[11] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose DeepFakes and face manipulations," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2019, pp. 83–92.

[12] J. Yang, S. Xiao, A. Li, G. Lan, and H. Wang, "Detecting fake images by identifying potential texture difference," *Future Gener. Comput. Syst.*, vol. 125, pp. 127–135, Dec. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X21002387

[13] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[14] Z. Yu *et al.*, "Searching central difference convolutional networks for face anti-spoofing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5295–5305.

[15] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," Jun. 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[16] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2889–2898.

[17] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 3207–3216.

[18] B. Dolhansky *et al.*, "The DeepFake detection challenge (DFDC) dataset," Jun. 2020, *arXiv:2006.07397*. [Online]. Available: http://arxiv.org/abs/2006.07397

[19] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, Jul. 2019.

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[21] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," Dec. 2013, *arXiv:1312.6114*. [Online]. Available: https://arxiv.org/abs/1312.6114

[22] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[23] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[24] J. A. Redi, W. Taktak, and J.-L. Dugelay, "Digital image forensics: A booklet for beginners," *Multimedia Tools Appl.*, vol. 51, no. 1, pp. 133–162, Jan. 2011.

[25] J. Lukäš, J. Fridrich, and M. Goljan, "Digital camera identification from sensor pattern noise," *IEEE Trans. Inf. Forensics Security*, vol. 1, no. 2, pp. 205–214, Jun. 2006.

[26] D. Cozzolino, G. Poggi, and L. Verdoliva, "Splicebuster: A new blind image splicing detector," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Nov. 2015, pp. 1–6.

[27] B. Bayar and M. C. Stamm, "Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2691–2706, Nov. 2018.

[28] F. Matern, C. Riess, and M. Stamminger, "Gradient-based illumination description for image forgery detection," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1303–1317, 2020.

[29] T. H. Thai, R. Cogranne, F. Retraint, and T.-N.-C. Doan, "JPEG quantization step estimation and its applications to digital image forensics," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 1, pp. 123–133, Jan. 2017.

[30] T. Bianchi, A. Piva, and F. Perez-Gonzalez, "Near optimal detection of quantized signals and application to JPEG forensics," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Nov. 2013, pp. 168–173.

[31] M. Kirchner, "Linear row and column predictors for the analysis of resized images," in *Proc. 12th ACM Workshop Multimedia Secur. (MM&Sec)*, 2010, pp. 13–18.

[32] J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 868–882, Jun. 2012.

[33] M. Huh, A. Liu, A. Owens, and A. A. Efros, "Fighting fake news: Image splice detection via learned self-consistency," in *Computer Vision—ECCV*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 106–124.

[34] O. Mayer and M. C. Stamm, "Learned forensic source similarity for unknown camera models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2012–2016.

[35] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1053–1061.

[36] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Distinguishing computer graphics from natural images using convolution neural networks," in *Proc. Inf. Forensics Secur.*, 2018, pp. 1–6.

[37] R. Wang *et al.*, "FakeSpotter: A simple yet robust baseline for spotting AI-synthesized fake faces," in *Proc. 29th Int. Joint Conf. Artif. Intell. (IJCAI)*, C. Bessiere, Ed., Jul. 2020, pp. 3444–3451.

[38] Y. Li, M.-C. Chang, and S. Lyu, "In Ictu oculi: Exposing AI created fake videos by detecting eye blinking," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2018, pp. 1–7.

[39] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2018, pp. 1–7.

[40] L. Li *et al.*, "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5001–5010.

[41] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "DeepFake video detection through optical flow based CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–3.

[42] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper, "Unmasking DeepFakes with simple features," Nov. 2019, *arXiv:1911.00686*. [Online]. Available: http://arxiv.org/abs/1911.00686

[43] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," Jul. 2020, *arXiv:2007.09355*. [Online]. Available: http://arxiv.org/abs/2007.09355

[44] R. Tolosana, S. Romero-Tapiador, J. Fierrez, and R. Vera-Rodriguez, "DeepFakes evolution: Analysis of facial regions and fake detection performance," Apr. 2020, *arXiv:2004.07532*. [Online]. Available: http://arxiv.org/abs/2004.07532

[45] Z. Liu, X. Qi, and P. H. S. Torr, "Global texture enhancement for fake face detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8057–8066.

[46] A. Kumar, A. Bhavsar, and R. Verma, "Detecting DeepFakes with metric learning," in *Proc. 8th Int. Workshop Biometrics Forensics (IWBF)*, Apr. 2020, pp. 1–6.

[47] J. Hernandez-Ortega, R. Tolosana, J. Fierrez, and A. Morales, "DeepFakesON-phys: DeepFakes detection based on heart rate estimation," Oct. 2020, *arXiv:2010.00400*. [Online]. Available: http://arxiv.org/abs/2010.00400

[48] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of synthetic portrait videos using biological signals," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 15, 2020, doi: 10.1109/TPAMI.2020.3009287.

[49] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[51] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[52] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," Mar. 2016, *arXiv:1603.07285*. [Online]. Available: http://arxiv.org/abs/1603.07285

[53] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. ICLR*, 2016, pp. 1–13.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[55] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–16.

[56] K.-Y. Lin and G. Wang, "Hallucinated-IQA: No-reference image quality assessment via adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 732–741.

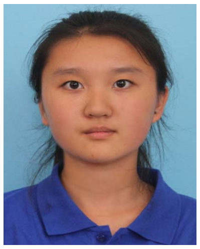
[57] J. Yang *et al.*, "No reference quality assessment for screen content images using stacked autoencoders in pictorial and textual regions," *IEEE Trans. Cybern.*, early access, Oct. 15, 2020, doi: 10.1109/TCYB.2020.3024627.

[58] Y. Li and S. Lyu, "Exposing DeepFake videos by detecting face warping artifacts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 46–52.

[59] L. Bondi, E. Daniele Cannas, P. Bestagini, and S. Tubaro, "Training strategies and data augmentations in CNN-based DeepFake video detection," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2020, pp. 1–6.

[60] N. Bonettini, E. Daniele Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of CNNs," Apr. 2020, *arXiv:2004.07676*. [Online]. Available: http://arxiv.org/abs/2004.07676

Jiachen Yang (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in communication and information engineering from Tianjin University, Tianjin, China, in 2002, 2005, and 2009, respectively. From 2014 to 2015, he was a Visiting Scholar with the Department of Computer Science, School of Science, Loughborough University, U.K. In 2019, he was a Visiting Scholar with Embry-Riddle Aeronautical University. He is currently a Professor with the School of Electrical and Information Engineering, Tianjin University, where he is the Leader of the Laboratory of Stereo Visual Information Processing. His research interests include image processing, artificial intelligence, and information security. In these areas, he has published more than 100 technical articles in refereed journals and proceedings, including IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, and IEEE TRANSACTIONS ON BROADCASTING. He is also on the editorial boards of IEEE ACCESS, *Sensors*, and *Multimedia Tools and Applications*, and held special issue on IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE ACCESS, and *Sensors*, as a Lead Guest Editor.



Aiyun Li received the B.S. degree in communication and information engineering from Tianjin University, Tianjin, China, in 2019, where she is currently pursuing the M.S. degree with the School of Electrical and Information Engineering. Her research interest includes digital image forensics.

Shuai Xiao is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering, Tianjin University, China. His research interests are machine learning, computer vision, and pattern recognition.

Wen Lu (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, China, in 2002, 2006, and 2009, respectively. Since 2009, he has been with the School of Electronic Engineering, Xidian University, where he is currently a Professor. From 2010 to 2012, he was a Post-Doctoral Research Fellow with the Department of Electronic Engineering, Stanford University, Stanford, CA, USA. He has authored or coauthored more than two books and around 50 technical articles in refereed journals and proceedings, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, *Information Sciences*, and *Neurocomputing*. His current research interests include multimedia analysis, computer vision, pattern recognition, and deep learning. He is on the editorial boards and serves as a Reviewer for many journals, such as IEEE TRANSACTIONS ON IMAGE PROCESSING and IEEE TRANSACTIONS ON MULTIMEDIA.

Xinbo Gao (Senior Member, IEEE) received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a Research Fellow with the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Post-Doctoral Research Fellow with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. He is with the Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications. He is currently a Cheung Kong Professor of the Ministry of Education, China, a Professor of *Pattern Recognition and Intelligent System*, and the Director of the State Key Laboratory of Integrated Services Networks, Xi'an. He has published six books and around 300 technical articles in refereed journals and proceedings. His current research interests include multimedia analysis, computer vision, pattern recognition, machine learning, and wireless communications. He is a Fellow of the Institute of Engineering and Technology and Chinese Institute of Electronics. He served as the General Chair/Co-Chair, the Program Committee Chair/Co-Chair, or a PC Member for around 30 major international conferences. He is also on the editorial boards of several journals, including *Signal Processing* (Elsevier) and *Neurocomputing* (Elsevier).