Spatial–Temporal Frequency Forgery Clue for Video Forgery Detection in VIS and NIR Scenario

Yukai Wang, Chunlei Peng, Member, IEEE, Decheng Liu, Nannan Wang, Member, IEEE, and Xinbo Gao, Senior Member, IEEE

Abstract—In recent years, with the rapid development of face editing and generation, more and more fake videos are circulating on social media, which has caused extreme public concerns. Existing face forgery detection methods based on frequency domain find that the GAN forged images have obvious grid-like visual artifacts in the frequency spectrum. But for synthesized videos, these methods only confine to a single frame and pay little attention to the most discriminative part and temporal frequency clue among different frames. To take full advantage of the rich information in video sequences, this paper performs video forgery detection on both spatial and temporal frequency domains and proposes a Discrete Cosine Transform-based Forgery Clue Augmentation Network (FCAN-DCT) to achieve a more comprehensive spatial-temporal feature representation. FCAN-DCT totally consists of a backbone network and two branches: Compact Feature Extraction (CFF) module and Frequency Temporal Attention (FTA) module. We conduct thorough experimental assessments on three visible light (VIS) based datasets (i.e., FaceForensics++, Celeb-DF (v2), WildDeepfake), and our self-built video forgery dataset DeepfakeNIR, which is the first video forgery dataset on near-infrared (NIR) modality. The experimental results demonstrate the effectiveness and robustness of our method for detecting forgery videos in both VIS and NIR scenarios. DeepfakeNIR and code are available at https://github.com/AEP-WYK/DeepfakeNIR.

Index Terms—Near-infrared face, face forgery detection, deepfake.

I. INTRODUCTION

IN RECENT years, with the widespread application of Deep Neural Network (DNN) and Generative Adversarial Networks (GAN) technologies, face generation, and editing technology have become more and more realistic and controllable. However, the abuse of this technology has caused widespread public concern. The current mainstream face forgery technologies include face replacement [1], [2], expression replay [3], [4], attribute editing [5], [6], and non-target face generation [7], [8]. With the rapid development of these tampering approaches, it is extremely hard to distinguish the real and fake videos on social media only with naked eyes. Thus, the malicious spread of these fake videos on the Internet has aroused great challenges for society and citizens. On the other hand, near-infrared (NIR) technology has been widely used in heterogeneous face recognition systems. Compared to traditional visible light (VIS) face recognition, NIR face recognition is an effective method for overcoming the impact of illumination changes and low light conditions. It uses a special purpose imaging capture hardware, in which active NIR lamps installed around the camera lens illuminate the face from the front, and the NIR camera captures the front-light NIR facial image. The NIR approach usually achieves significantly higher performance than the VIS approach. In the real-world scenario, NIR technology is widely used in face unlocking applications and scene monitoring. For example, iPhone X uses Face ID face unlocking function and uses a near-infrared camera to solve the lighting problem when unlocking. At the same time, near-infrared technology is also widely used in iris unlocking. The iris is a ring-shaped film located between the black pupil and the white sclera on the surface of the human eyes. Under the modality of infrared light, it can present rich visual features such as spots, stripes, filaments, coronal, crypt, etc. However, due to the huge gap such as the difference in the collected bands and colors between the VIS and NIR modality, many related works found that the pretrain

1https://www.pocket-lint.com/phones/news/apple/142207-what-is-apple-face-id-and-how-does-it-work
2https://www.csoonline.com/article/3197933/hackers-easily-trick-iris-scanner-to-unlock-samsung-galaxy-s8.html
network under the VIS modality performs poor accuracy in the heterogeneous face recognition (HFR) scenario. Reference [9] pointed out that due to the difference in spectral components between NIR and VIS modalities, visible light and near-infrared (VIS-NIR) face recognition is still a challenging task. Furthermore, the NIR-based face recognition systems are also threatened by face forgery attacks. An attacker can inject a fake near-infrared face into the NIR-based access control system to escape identity authentication and thus avoid tracking; the NIR-based face unlocking technology used in the Xiaomi 8 phone was successfully cracked by simply using a black and white printer printout and adding shadows around the eyes and nose of the face to give it a more three-dimensional appearance. In addition, near-infrared is often used for night video surveillance because it is robust to illumination. The application in NIR scenario are shown in Fig. 1. However, NIR face recognition and scene monitoring also suffer from security issues of forgery attacks. In NIR face recognition applications, attackers may use forged face images to deceive the NIR face recognition system. The face recognition system compares forgery faces with those in the gallery and treats them as legitimate one. Thus, they could exploit this vulnerability to gain access to protected resources or systems by using forgery facial images. In NIR monitoring applications, due to the poor light conditions in the environment, a surveillance video in the NIR modality will cause misunderstandings for viewers, thus playing a misleading role. For example, a fake video may make people think that something happened when it didn’t, causing the people in the fake video to suffer from public opinion. Considering the wide application of authentication under the NIR scenario, the severe challenges of face forgery and detection in the NIR modality also should be considered [10].

To address these challenges, a lot of research efforts have been dedicated to the field of face forgery detection. Approaches for face forgery detection can be broadly divided into three categories: image/frame-based, frequency-based and video-based methods. Reference [11] first proposed the GAN fingerprint phenomenon, it found that images generated by different GAN models leave diverse features which are called fingerprint features. Inspired by the principle that synthesized images in the frequency domains are often more able to reflect the visual artifacts as shown in Fig. 2, some attempt to obtain information from frequency domains, such as Discrete Fourier Transform (DFT) and Discrete Cosine Transform (DCT). Reference [12] proposed a novel Progressive Enhancement Learning (PEL) framework to exploit fine-grained frequency information and fully decouple tampering traces in the frequency domain. Recently, more and more researchers believe that the image/frame-based method ignores the temporal inconsistencies between frames. Therefore, they directly extract video-level features for forgery detection. MesoNet [13] proposed a compact facial video forgery detection network to automatically detect face manipulations in videos. Reference [14] proposed a two-stream detection network on a spatial and temporal level that achieved significant improvements on post-processing compressed low-pixel videos. However, existing methods did not take both the spatial and temporal frequency clue into account which restricts the performance and robustness of video forgery detection.

Inspired by [15], we also visualize the motion of vertical slices at intermediate horizontal positions in both VIS and NIR videos. It is obvious that the motion in the real video is smoother than the fake one under the VIS modality, while this phenomenon is not obvious under the NIR modality. At the same time, we also visualize the temporal clues in the frequency domain. It can be seen that whether under the VIS or NIR modality, the forged video loses too many high-frequency components and lose the randomness of the frequency component distribution, which shows that the temporal inconsistency in the frequency domain may serve as an effective indicator for forgery video detection.
slight inconsistency in NIR modality, it is not very obvious. However, the temporal forgery clue is more obvious in the frequency domain regardless of it is based on the VIS or NIR modality. We can see that fake videos often lack more high-frequency components, and there is no random distribution of high and low frequencies between frames like real video spectrum. This may be because most forgers often perform operations such as blurring and compression at the junction in the process of face replacement in order to make the face fusion more natural. All frames in the video use the same post-processing parameters causing the problem of consistent high and low frequency components in the frequency domain. Therefore, the temporal inconsistency in the frequency domain may serve as an effective indicator for forgery video detection. Thus, based on our discovery, we aim to propose a Discrete Cosine Transform-based Forgery Clue Augmentation Network (FCAN-DCT) for video forgery detection to take both the spectrum spatial and temporal forgery clue into account and further achieve a more comprehensive representation of fake videos. Compared with image/frame-based methods, our method additionally introduces temporal information through attention mechanism. Compared with the frequency-based method, our method adopts a block strategy in the spatial domain by analyzing the discreteness of DCT, which greatly reduces the amount of computation. And our method creatively introduced the temporal forgery clue in the frequency domain. Compared with video-based methods, our method introduces an attention mechanism on multi-frame spectrums, which makes the exposure of forgery clues more obvious, and makes the extracted feature more representative and discriminative. Specifically, FCAN-DCT consists of a backbone network and two branches: Compact Feature Extraction (CFE) module and Frequency Temporal Attention (FTA) module. The backbone network first extracts the feature map of the video sequences and then converts the feature map of the RGB space to the frequency domain based on Discrete Cosine Transform (DCT) and the converted feature maps can be considered as frequency spectrums. CFE adopts a block strategy and uses the maximum value in each block to represent the weight of each of the frequency domain components. Therefore, after the frequency division, we achieve the set of compact and effective spectrum spatio-temporal information.

Moreover, to further explore and highlight the temporal frequency forgery clue, we design a Frequency Temporal Attention Module (FTA) to explicitly model the temporal frequency clue between different frames and build an attention map with channel attention mechanism. Therefore, with the compact and effective spatio-temporal information set and the 2D attention map, we can pay attention to the most discriminative features among a stack of frame sequences.

In addition, face forgery detection in multimodal scenarios is rarely explored except for [10] which is a NIR image forgery dataset. Therefore, we further construct a NIR video forgery dataset, DeepfakeNIR, to evaluate the robustness and generalization ability of our method and promote the research on video forgery detection in multimodal scenarios. Because there is no video dataset and forgery detection method specifically for NIR modality, we only used a general forgery method (i.e., deepfacelab [16]) similar to most of the VIS forgery datasets. In summary, the main contributions are as follows:

- We focus on the spectrum spatial-temporal frequency forgery clue for video forgery detection by paying attention to the discriminative features among frame sequences, which fills the gap of video forgery detection in the frequency domain.
- We propose a Discrete Cosine Transform-based Forgery Clue Augmentation Network (FCAN-DCT). The CFE and FTA modules make full use of the information in different video frames. These two branches are combined to achieve a more comprehensive spatial-temporal feature representation in the frequency domain.
- We firstly perform video forgery detection experiments on both VIS and NIR scenarios, including three wildly used three datasets (i.e., FaceForensics++, Celeb-DF (v2), WildDeepfake), and our newly constructed DeepfakeNIR dataset, to evaluate the effectiveness and robustness of the proposed method.
- We present a novel video forgery detection dataset (DeepfakeNIR) with diverse distributions in terms of posture, occlusion, and expression. There are a total of 3,847 NIR videos, where the ratio of real and fake NIR videos is near 1: 1. The proposed dataset can help facilitate the research of video forgery detection in heterogeneous modalities.

II. RELATED WORK

A. Facial Image Forgery Detection

Early face generation technologies often directly show obvious visual artifacts and inconsistencies in the facial area due to their uncontrollable characteristics. Many previous works employ face or head statistical inconsistencies for image forgery detection. For example, the inconsistency in estimating 3D head pose from face images was introduced in [17] to determine the authenticity of an image. Face X-ray [18] proposed a novel method aiming to predict the blending boundaries in fake video frames. Through the unremitting efforts of these works, facial image forgery detection has achieved remarkable success in different aspects. Reference [19] proposed a GAN Fingerprint Disentangling Network (GFD-Net) to disentangle the fingerprint feature from GAN-generated images. Reference [20] proposed a reconstruction-classification learning (RECCE) framework that better generalizes to unknown forgery patterns. Reference [21] showed good generalization ability on unseen datasets through intra-consistency within classes and inter-diversity between classes. Reference [22] tracked potential texture traces during image generation and showed good performance on most benchmarks. Reference [23] employed a strategy of adversarial learning to add adversarial perturbations to the regions that most forgery detectors on GAN-generated images interest. The results demonstrate the vulnerability of the current image-based forgery detection method. Aiming at these post-processed videos, [24] found that both the luminance and chrominance components in the image play an important role in the final detection result. This method achieves better performance in the RGB and YCbCr.
color spaces. The Artifacts-Disentangled Adversarial Learning (ADAL) framework proposed in [25] exhibited strong generalization ability and generality by disentangling the artifacts from irrelevant information. However, with the development of DeepFake generation techniques, these image-based detectors may fail to capture temporal inconsistencies across multiple frames.

B. Facial Video Forgery Detection

The early forgery generation technology is relatively limited, and it may show obvious temporal inconsistency between frames visually. Physiological signal based on blink frequency which is not well represented in synthetic fake videos is explored in [26]. Agarwal et al. [27] found that the tampering with the face area in the video can cause the pattern of facial expressions and head movements inconsistent with the person’s identity when speaking. Reference [15] proposed a novel temporal inconsistency modeling paradigm (TIM) for detecting face forgeries. However, these approaches are not universal. Recently, face generation and editing techniques are becoming more realistic and controllable, resulting in inconsistency artifacts in fake videos that are increasingly difficult to detect. Therefore, more related works begin to explore from multiple perspectives. Reference [28] proposed a deepfake video detection method based on the dissimilarity between audio and video sequences. Reference [29] proposed a novel method to justify the authenticity of videos by detecting the consistency of the emotions predicted by the speaker’s face and voice. Reference [14] proposed a spatial and temporal level two-stream detection network that achieved significant improvements on post-processing compressed low-pixel videos. In addition, more and more works begin to focus on the generalization of forgery detection tasks on unseen datasets. Reference [30] proposed a temporal convolution network that keeps the size of the temporal convolution kernel unchanged and reduces the size of the spatial convolution kernel to 1, making it more effective to extract temporal features and improve the generalization ability of the model. Reference [31] proposed an efficient and robust framework named LRNet for detecting deepfakes videos by temporal modeling of precise geometric features. The Local Temporal-aware Transformer-based Deepfake Detection (LTTD) method [32] employed a local-to-global learning approach that emphasizes the important temporal information within local sequences. Specifically, they introduced a Local Sequence Transformer (LST) that captures the temporal coherence of sequences within restricted spatial regions. Reference [33] observed that the process of creating deepfake videos often disrupts the statistical regularity that is present in genuine videos. Inspired by this phenomenon, it proposed a method for improving the generalization of deepfake detection by identifying the “regularity disruption” that does not occur in real videos. Reference [34] revealed that fake face videos exhibit truncation in the consecutive frames of optical flow imaging following dense optical flow processing. The forgery detection method proposed by [35] was able to capture local spatio-temporal relations and inconsistencies in deepfake videos, while existing sequence encoders are not sensitive to these features. Reference [36] proposed the use of self-blended images (SBIs) as a new type of synthetic training data for detecting deepfake videos. Reference [37] proposed a new way to spot deepfake images by looking for inconsistencies in the source features within the manipulated images. Inspired by the principle that the channel and spatial features of a robust forgery detector should be consistent in the temporal domain. Reference [38] introduced a commonality learning strategy for detecting forgeries in forgery videos. The goal of this approach is to improve the generalization performance of the forgery detector by learning common forgery features from various forgery databases.

C. Frequency-Based Forgery Detection

Considering that synthesized images in the frequency domain are more able to reflect the visual artifacts, a lot of efforts have been dedicated to the field of forgery detection in the frequency domain. Reference [39] conducted DCT operation on the image, and then inversely transforms back
to the spatial domain based on the low-frequency, mid-frequency and high-frequency information to obtain the spatial components of different frequency bands, which were finally sent to CNN to extract features. Reference [40] designed a sliding window DCT_IDCT module to improve the network performance by fusing the frequency information into the neural network. SPSL [41] leveraged both spatial image and phase spectrum information to effectively identify up-sampling artifacts in forged facial images, leading to improved transferability in detection performance. Reference [42] presented a Frequency-aware discriminative feature learning (FDFL) framework for detecting face forgeries. Reference [43] utilizes communal features in multiple frames across different domains (RGB and frequency patterns) to promote stability and consistency can adaptively adjust discriminative centers based on individual instance features, further improving detection performance. Reference [44] proposed a symmetric transformer model by adopting newly-designed attention-based strategies for channel variance and spatial gradients as the vital features, which greatly improves the robustness of deepfake video detection. Reference [45] proposed an adaptive frequency learning approach in the two-branch detection framework for face forgery detection. The approach adaptively decomposes frequency information with soft masks optimized using triplet loss and incorporates frequency characteristics into spatial clues with an attention module. Reference [46] proposed a novel approach that utilizes frequency attention and multi-view based knowledge distillation to improve the detection of low-quality compressed deepfake images. [47] proposes a novel approach to detecting face manipulation that utilizes a dual-branch network architecture and frequency-based feature refinement to improve generalization and robustness. All the aforementioned frequency-based methods pay little attention to the temporal inconsistency (as shown in Fig. 3) between the spectrum of different frames in real and fake videos directly. Our proposed FCAN-DCT not only explores the spatial frequency feature but also augments the temporal inconsistency clue in the frequency domain with the attention mechanism. Thus, the final spectrum spatial-temporal feature learned by our method can achieve a more comprehensive representation.

### III. Proposed Approach

Aiming at filling the gap of analyzing the entire video sequence instead of a single image in the frequency domain, we introduce the Discrete Cosine Transform-based Forgery Clue Augmentation Network (FCAN-DCT) for forgery detection. FCAN-DCT consists of a Compact Feature Extraction (CFE) module in section III-B and a Frequency Temporal Attention (FTA) module in section III-C. Fig. 4 shows the framework of FCAN-DCT.

#### A. Preliminaries

Since the tampering area in the forgery task is only limited to the face area, in order to prevent other areas from affecting the final detection result, we first use the Single Shot Scale-invariant Face Detector (SS3-FD) [48] to detect and extract the face area in each frame. Then, we uniformly sample $N$ cropped frames as the video input sequence of the model $I = \{f_1, f_2, f_3, \ldots, f_N\}$. Each selected frame is input to the backbone network and finally outputs the feature map representation of each frame $M = \{m_1, m_2, m_3 \ldots m_N\}$. We transform these resulting feature maps into the frequency domain using the Discrete Cosine Transform (DCT). Much like the Discrete Fourier Transform (DFT), DCT represents a sequence of points in space as the sum of cosine functions at different frequencies. Given a pixel matrix in RGB space $f \in \mathbb{R}^{H \times W}$, the 2D-DCT is defined as:

$$
D(u, v) = c(u) c(v) \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} f(i, j) \cos \left[ \frac{(i + 0.5) \pi}{H} u \right] \cos \left[ \frac{(j + 0.5) \pi}{W} v \right]
$$

where $u$, $v$ are the coordinates of the DCT spectrum matrix $D(u,v) \in \mathbb{R}^{H \times W}$. Compared with DFT, DCT has better energy compaction performance, and can directly crop those unimportant frequency domain regions and represent a signal or image with fewer coefficients while preserving more information, so it is widely used in image compression tasks, such as jpeg compression. In recent years, with the rapid development of deep learning, the operation of DCT also attracted considerable interest in forgery detection tasks.

#### B. Compact Feature Extraction Module

We adopt Compact Feature Extraction (CFE) module to achieve the set of compact and effective spectrum spatio-temporal information. CFE is mainly divided into two steps: 1) Use the 2D-DCT in equation (1) to transform the input pixel feature map $M = \{m_1, m_2, m_3 \ldots m_N\}$ to the DCT frequency domain $D = \{d_1, d_2, d_3, \ldots d_N\}$, while retaining more important information and discarding unimportant information. At the same time, artifacts that appear to be fake images are shown. 2) Compress the obtained frequency domain sequence to obtain a more compact feature representation. Inspired by [49], [50], and [51], which suggest that the components present in the spectrum of medium and high-frequency bands exhibit significant inconsistencies between real and synthesized images, we designed a weight matrix to enhance the spatial forgery clues in the spectrum of these bands. The resulting product of multiplying the weight matrix with the DCT spectrum serves as the input to the CFE module. Here, $\beta = \sqrt{2}$. For 2D DCT, the equation can be expressed as:

$$
F(u, v) = \beta^\alpha D(u, v),
\begin{cases}
\alpha = 0, & 0 \leq u + v < \frac{H}{3} \\
\alpha = 1, & \frac{H}{3} \leq u + v \leq \frac{2H}{3} \\
\alpha = 2, & \frac{2H}{3} < u + v < H
\end{cases}
$$
where \( u, v \) are the coordinates of the DCT spectrum, and \( H \) is the height of the spectrum. The visualization of weight matrix are shown in Fig. 5.

After multiplication with the coefficient matrix, the resulting spectrum sequence is \( F \in \mathbb{R}^{N \times C \times H \times W} \). It should be noted that applying self-attention mechanism on the spectrum spatial domain would incur significantly higher time and space complexity. Therefore, we manually designed a weight matrix based on prior knowledge to enhance the performance. Considering the discrete nature of DCT transform, that is, some of the most valuable and effective information is often distributed across different parts of the spectrum, we utilize a block-based approach to extract the spatio-temporal information from the spectrum. Specially, we divide the spectrum \( F \in \mathbb{R}^{N \times C \times H \times W} \) into \( K \) frequency bands in the spatial dimensions of \( H \) and \( W \), resulting in a set of frequency bands \( F = \{F_1, F_2, F_3 \ldots F_K\} \). For each frequency band, we use the maximum response to represent the most important information within the block. By doing so, we ensure that the model focuses on the critical regions during the training process. After the frequency division in the CFE module, we obtain a stack of compact spectrum spatial feature maps for the video sequence. To this end, we consider \( F_{CFE} \in \mathbb{R}^{N \times C \times K} \) as the set of effective spectrum spatio-temporal information.

### C. Frequency Temporal Attention Module

To further explore the temporal frequency clue among multiple frames, we design a Frequency Temporal Attention (FTA) module. Similar to CFE, the input of the FTA module is also the DCT spectrum \( F \in \mathbb{R}^{N \times C \times H \times W} \). The goal of this module is to construct an attention map \( A \in \mathbb{R}^{N \times H \times W} \) based on the DCT frequency spectrum. First, we conduct the \( L2 \) normalization in the channel dimension.

\[
A(n, h, w) = \left( \sum_{i=0}^{C-1} (F_{n,i,h,w})^2 \right)^{-\frac{1}{2}}
\]

(4)

where \( C \) is the number of channels. \( A(n, h, w) \) represents the attention score of the \( n^{th} \) frame spectral feature map at position \((h, w)\). Then in order to control all the values in the attention matrix to be between 0-1, the results are normalized as:

\[
A'(n, h, w) = \frac{A(n, h, w)}{\sum_{i=0}^{H-1} \sum_{j=0}^{W-1} A(n, i, j)}
\]

(5)

Following the same principle in CFE, we also divide the feature map matrix into \( K \) blocks, and for each region, we use the sum of them as the attention score. The results are:

\[
A''(n) = \{A''(n, 1), A''(n, 2) \ldots A''(n, K)\}
\]

(6)

Because after performing \( L2 \) norm operation to the DCT spectrum \( F \), which makes the values positive and makes the attention map have a sum of 1 in the \( T \) dimension, we adopted \( L1 \) norm based on \( A'' \in \mathbb{R}^{N \times K} \), and further explore the temporal relationship between frame sequences. The 2D attention matrix is expressed as:

\[
A_{FTA}(n, k) = \frac{A''(n, k)}{\sum_{i=0}^{N-1} A''(i, k)}
\]

(7)

In the end, to integrate both spatial and temporal forgery clues, we combine the compact and effective spatio-temporal information set \( F_{CFE} \in \mathbb{R}^{N \times C \times K} \) with the frequency temporal attention map \( A_{FTA} \in \mathbb{R}^{N \times K} \). We achieve this fusion by multiplying the elements in the \( N \) and \( K \) dimensions, and then summing the results. By doing so, we obtain a comprehensive representation of the video sequence that incorporates both spatial and temporal clues. This approach allows us to effectively combine information from spatial and temporal domain of the spectrum and produces a more robust representation of the video. The final spectrum spatial-temporal frequency forgery feature \( f_c \in \mathbb{R}^C \) is constructed as:

\[
f_c = \sum_{n=0}^{N-1} \sum_{k=0}^{K-1} F_{CFE}(n, c, k) \times A_{FTA}(n, k)
\]

(8)

Therefore, we can pay attention to both the spatial and temporal frequency forgery clue among a stack of frame sequences.

### IV. Experiments and Results

In this section, we first introduce the benchmarks and the parameter settings in the experiments. Furthermore, we conduct sufficient experiments to demonstrate the effectiveness of our method FCAN-DCT. The results on the self-built near-infrared modality dataset DeepfakeNIR for face forgery detection also demonstrate that FCAN-DCT has good generalization across NIR modalities.

#### A. Experimental Settings

**Datasets:** In this paper, we use four video-based datasets for forgery detection: WildDeepfake [52], Celeb-DF (v2) [53], FaceForensics++ [54] and our newly constructed DeepfakeNIR datasets.

- **Celeb-DF (v2)** is a large-scale dataset proposed on the basis of v1, which contains 590 original real videos collected from YouTube and 5,639 corresponding fake videos with diverse distribution in terms of gender, age, and ethnic group.
Fig. 6. We conduct experiments on four datasets to verify the effectiveness of FCAN-DCT. The first three benchmarks are common datasets based on the visible light (VIS) modality: (a) FaceForensics++, (b) Celeb-DF (v2), (c) WildDeepfake. The other is our proposed video dataset based on near-infrared (NIR) modality: (d) DeepfakeNIR.

- **FaceForensics++** is a large-scale benchmark dataset which contains 1,000 original videos from YouTube and 4,000 fake videos generated by four typical manipulation methods, i.e., Deepfakes (DF), Face2Face (F2F), FaceSwap (FS) and NeuralTextures (NT). Each method generates 1,000 fake videos corresponding to the original video. There are three versions in terms of compression level, i.e., raw, lightly compressed (HQ), and heavily compressed (LQ).

- **WildDeepfake** contains 7,314 facial sequences extracted from 707 Deepfake videos. The videos are all collected from bilibili and YouTube, and their face-swapping videos are synthesized by various methods, so the detection is more difficult.

- **DeepfakeNIR**: In addition to the visible light (VIS) modality, we also experimented with our method in the near-infrared (NIR) modality. DeepfakeNIR contains 3,847 videos in total. Specifically, the detailed construction process is as follows: first, we divide the 59 videos of NIR videos collected from Near Infrared Face Database [55] into six groups, (e.g. 1-10, 11-20, 21-30, 31-40, 41-50, 51-60). It is worth noting that since the author did not provide the 15th video, we get a total of 59 videos; Then, we use the 10 video identities of the former group to replace the corresponding videos in the latter using deepfacelab tool [16], and we get a total of 58 fake videos; Eventually, we divided these videos into 1,939 real and 1,908 fake videos in terms of posture, occultation, and expression. Furthermore, we apply various perturbations such as local block-wise distortion (BW), white Gaussian noise in color (GNC), color contrast change (CC), gaussian blur (GB) and JPEG compression (JPEG), etc. to better mimic videos in real-world scenarios. Specifically, we divide each of these perturbations into five intensity levels. Then, we randomly select the five types of perturbation and intensity with equal probability and finally generate corresponding 1,939 real and 1,908 fake perturbated videos. The ratio of five perturbation types applied in the dataset is roughly 1:1:1:1:1. In addition, similar to the FF++ dataset, we applied H.264 codec compression on both real and fake videos in DeepfakeNIR with c23 compression levels which represents videos compressed with a CRF setting of 23. Example frames in [55] and our DeepfakeNIR are shown in Fig. 7.

1) **Preprocessing**: We select 16 frames from each video, then for each frame, we extract facial key points and crop out the head region of the person through the Single Shot Scale-invariant Face Detector (S3F D) [48]. Examples are shown in Fig. 6. Finally, we normalize all the faces with mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225], and resize them to a fixed size which is 256 x 256 for ResNet50 [56] and 299 x 299 for Xception [57]. **Training**: In our experiments, we use ResNet50 [56] or Xception [57] as the final backbone of our approach. For retaining more information on the video sequence, we just remove the last average pooling and fully connected layer. We optimize the networks with Adam optimizer with the beta1 = 0.9 and beta2 = 0.999. The start learning rate lr = 0.0001 and it drops by 10 every time the accuracy does not increase after 7 consecutive epochs. The evaluation metrics in our experiments are accuracy (ACC) and Area Under Curve (AUC).

2) **Selection of division number K and the number of frames N in each video**: We conducted experiments on four types of synthetic media methods (i.e., Deepfakes, Face2Face, FaceSwap, and NeuralTextures) on FF++, Celeb-DF (v2) and WildDeepfake, and used the average AUC results across the three datasets as the major metric to determine the values of K and N. Experimental results are shown in Table I. To provide a more intuitive representation of the results, we plotted the results for experiments with K = 2 × 2, 4 × 4, 8 × 8, and N = 4, 8, 16, 32 in Fig. 8. On the one hand, it can be found that no matter how the parameter K is selected, the curve shows an obvious upward trend when N is less than 16. However, when N exceeds 16, the average AUC result drops extremely. On the other hand, we find that no matter how the parameter
methods demonstrate the effectiveness of expressing features in the frequency domain. Compared with F3-Net [39], the ACC and AUC of our proposed method increased by 5.69% and 6.21%. PEL achieves 84.14% ACC and 91.61% which is the SOTA frequency-based method on WildDeepfake, while our method improves the ACC and AUC by a large margin of 2.21% and 2.12%. Compared to the video-based method, i.e., STIL [15], the performance gains can be observed on both three benchmarks. Especially on WildDeepfake dataset, our method reaches a state-of-the-art result by improving the ACC by 2.23%. It should be noted that existing methods have achieved almost ideal results on the FF++ (HQ) dataset. Therefore, we also considered the low-quality (LQ) versions of the FF++ dataset and achieved remarkable progresses as shown in Table III. While, our method achieve a significant advantage on the harder and realistic WildDeepfake datasets that suffer from various compressions and other perturbations. The remarkable progress demonstrates the spectrum spatial-temporal feature extracted by FCAN-DCT is efficient and effective.

1) 3D-CNN method: Considering video-based face forgery detection tends to realize the inconsistency between frames, it seems more reasonable to use 3D-CNN directly considering the temporal relationship, we test some of the most popular 3D-CNN based networks and some SOTA video-based forgery detection method utilizing temporal features including RCN [63], R2Plus1D [64], I3D [65], MC3 [64] and R3D [66] on Celeb-DF (v2) [53]. Experimental results are shown in Table IV. It can be seen that FCAN-DCT outperforms all existing temporal-based networks to achieve competitive accuracy and AUC with state-of-the-arts. Especially when ResNet50 is selected as the backbone network, the accuracy and AUC reached 99.8% and 99.99% respectively which is a near-ideal result. R3D [66] is a 3D CNNs based on ResNet toward a better action representation. R2Plus1D [64] decomposes 3D convolutional filters into separate spatial and temporal components and improves the accuracy of action detection. However, few of these methods are specifically designed for video forgery detection and too many parameters of 3D-CNN models make it difficult to apply to lightweight scenarios. Moreover, these 3D-CNN-based detection works focus too much on time series modeling in DeepFake video detection, but ignoring the spatially imperceptible details, which limits the detection effect of DeepFake videos. Our method not only notices the most important part in different frames but also explores temporal information in the frequency domain and finally presents a comprehensive spatial-temporal representation. In addition, apart from the backbone network, no additional parameters are added. Therefore, our method is more capable and effective in capturing spatial-temporal forgery inconsistencies than 3D-CNN methods.

2) Cross-Dataset Evaluation: To demonstrate the effectiveness and generalizability of our proposed method, we conducted cross-dataset experiments on three popular deepfake datasets, namely FF++, Celeb-DF (v2), and WildDeepfake. Specifically, we evaluated the performance of our method trained on one dataset and tested on the other two datasets, and compared it with the state-of-the-art methods.
The experimental results show that our method achieves promising performance on all three datasets, indicating its ability to handle the diverse deepfake manipulations and generalize well to new datasets. We compare FCAN-DCT with several recently state-of-the-art methods including F3-Net [39], Face X-ray [18], LipForensics [67], HFF [68], CNNDetection [69], GANFingerprint [49], RECCE [20] and RFM [60].

Table V and VI show that our method can achieve much better generalization capability under cross-dataset scenarios. It can also be observed that the detection performance of most frequency-based methods, such as RECCE, F3-Net drops much more drastically under cross-dataset scenarios compared with that of the video-based methods (LipForensics). It is because these model confine to the static frequency artifacts leaved in the single spectrum but pay little attention to the inconsistency of multiple frames and thus the frequency-based methods based on these artifacts are incomplete and vulnerable. In the temporal domain, LipForensics proposes a novel lip forgery detection method that combines 3D convolutional neural networks and long short-term memory networks to
detect artificial tampering traces in lip videos. In the frequency domain, \( F^3 \text{Net} \) [39] learning feature representation by decomposing the frequency feature, thus is more sensitive to the forgery traces in the different frequency bands components. On the contrary, our model aims to explore both spectrum spatial artifacts and temporal frequency component irregularities, and exhibits better generalization performance.

We further train a classifier based on the high-quality version of FF++ which consist of four forgery methods (i.e., Deepfakes, Face2Face, FaceSwap and NeuralTextures). We select one of the forged methods to train and test on the other methods. The comparative results are shown in Table VI. It can be seen that our model performs competitive results in most cases. Face X-ray [18] achieves a relatively better generalization ability by detecting the blending evidence. HFF [68] captures the blending effects in the noise space by leveraging both textures and noises, therefore generalizing better from one method to another. Our model leverages both spatial and temporal information in frequency domain which confirms the effectiveness and robustness of our proposed FCAN-DCT.

C. Experimental Results in NIR Scenario

In this subsection, we conducted the experiments with state-of-the-art Deepfake detection methods such as CNNDetection [69], GANFingerprint [11] RECECE [20], RFM [60], HFF [68] and our proposed FCAN-DCT on our newly constructed DeepfakeNIR dataset, to evaluate the robustness of the proposed method. The intra and inter-dataset experimental results are shown in Table VII and VIII. In the intra-dataset evaluations, due to the widespread use of compressed videos on social media networks, we introduced image-based disturbances such as JPEG compression in our CFE module, and video compression in FTA, which is based on single-frame image spectrum spatial forgery clues, and video compression in FTA, which is based on whole-video spectrum temporal forgery clues. We conducted experiments on DeepfakeNIR with the type and strength of the
disturbances to be consistent with those in Deepforensics-1.0 [71] and FF++ [54]. The results in Table VII demonstrate that our method outperforms the approaches that only explore spatial frequency domain forgery clue (i.e., HFF [68] and GANFingerprint [49]) in both ACC and AUC. In the cross-modality experiments, we use three datasets under the VIS modality for training and testing on DeepfakeNIR. Our method achieved the best ACC and AUC metrics across all datasets, especially when training on Celeb-DF (v2) and WildDeepfake, our method outperformed other state-of-the-art methods by nearly 10%. The considerable performance demonstrates that the temporal frequency domain forgery clue introduced in our method exhibit good generalization.

D. Ablation Study

In this section, we conduct sufficient experiments to demonstrate the effectiveness of our method and analyze the impact of hyperparameter selection on the results. In addition, we perform ablation studies on the proposed FCAN-DCT. We train and test CFE and FCAN-DCT separately, and the experimental result of each model is shown in Table IX. It should be noticed that the temporal clue extracted in FTA is a weight attention map for the CFE module which is designed for extracting the spatial frequency feature. The final spatial-temporal representation can only be obtained with the joint cooperation of both two branches. Therefore, we just operate the CFE module in this section while it is not available to experiment only with the attention weights obtained by the FTA branch.

1) Effectiveness of the Coefficient Weight Matrix: Previous works [42], [49], [51] on forgery detection in the frequency domain for a single image believe that there exists a significant difference between the real and fake images in the midium and high frequencies, and this difference is more obvious as the frequency increases. Inspired by these works, we construct a coefficient weight matrix whose height and width are the same as those of the DCT spectrum. While keeping the low-frequency components fixed, we increase the weights of the midium and high-frequency components so that the model pays more attention to the area for fake clues. The experimental results are shown in Table IX. When there is no coefficient weight matrix, the ACC of merely applying the original DCT spectrum reaches 85.37% while adding the coefficient weight matrix in the frequency transform will increase the ACC and AUC by 0.95% and 0.27%, respectively. The progress demonstrates the coefficient weight matrix is efficient and effective.

2) Effectiveness of the Temporal Attention Map: By comparing with and without the temporal attention map settings in FCAN-DCT, it is clear that the results of FCAN-DCT outperform the CFE branch with only the spatial compact feature. For example, when there is no temporal frequency attention map, the ACC of merely applying the spectrum spatial forgery clue reaches 85.13%. Furthermore, adding the temporal attention map to the spectrum spatial clue will increase the ACC and AUC by 1.22% and 1.29%, respectively, which denotes that the temporal attention map learned by FTA is effective in video forgery detection.

3) Effectiveness of the Maximum Block Strategy: Due to the discretleness of DCT, the effective information embedded in DCT frequency spectrum is sparse. To make the frequency feature expression more compact, we divide the whole spectrum F into K frequency band, and then use the max response to represent each corresponding frequency band in the CFE module. We conduct the strategy comparison experiment by taking the maximum, minimum, and average value within the block on the WildDeepfake dataset. Our chosen framework backbone is ResNet50 and the setting of K = 4 × 4, N = 16.

TABLE IX
WE CONDUCTED ABLATION EXPERIMENTS ON WILDDEEPFAKE TO EXPLORE THE IMPACT OF EACH MODULE ON THE EXPERIMENTAL RESULTS. SPECIFICALLY, WE EXPLORED WEIGHT MATRIX FOR THE SPATIAL DOMIAN OF THE SPECTRUM, AS WELL AS TEMPORAL ATTENTION MAPS FOR THE TEMPORAL DOMAIN. THE BACKBONE APPLIED IN THE NETWORK IS ResNet50 AND THE SETTING OF K = 4 × 4, N = 16.

| Weight matrix | Temporal attention map | ACC (%) | AUC (%) |
|---------------|------------------------|---------|---------|
| -             | -                      | 83.88   | 92.45   |
| ✓             | -                      | 85.13   | 92.45   |
| -             | ✓                      | 85.37   | 93.47   |
| ✓             | ✓                      | 86.35   | 93.74   |

E. Visualizations

1) Grad-CAM: To better understand the effect of our spectrum spatial-temporal forgery clue on the original image, the visualization of Grad-CAM [72] on FF++ [54] is shown in Fig. 9. In terms of spatial dimension, our method is sensitive to forgery traces. In the temporal dimension, our method
TABLE X

| Method      | ACC (%) | AUC (%) |
|-------------|---------|---------|
| Min         | 83.75   | 91.85   |
| Avg         | 83.38   | 91.14   |
| Max (Ours)  | 86.35   | 93.74   |

Fig. 9. Grad-CAM visualization of real and four fake methods on FF+++, i.e., Real, Face2Face (F2F), FaceSwap (FS), Deepfakes (DF) and NeuturalTextures (NT). In the spatial dimension, our method focuses on different regions for different forgery methods, and in the temporal dimension, our method will focus more on the regions where obvious actions occur between frames. These are all benefited from our spectrum spatial-temporal forgery clue.

has a clear distinction on the heatmaps of real and fake images, where our method pays more attention to regions with obvious motion differences between the frame sequences. For instance, when there are significant movements of the mouth or eyes in the front and rear frames, our method will focus on these abnormal areas. This is reasonable because forgery clues tend to be exposed with greater probability in these regions. Moreover, heatmaps in FF+++ also show that our method differs in salient regions for various forgery techniques such as for NeturalTextures, our method pays more attention to the mouth region. Therefore, our proposed FCAN-DCT is not limited to a specific falsification pattern and thus improves the generalization ability.

2) T-SNE: In order to better demonstrate the effectiveness of our method for extracting video features, we have performed T-SNE feature visualization results on both VIS and NIR datasets, i.e., Celeb-DF (v2) [53] and DeepfakeNIR, respectively, as shown in Fig. 10. It can be seen that video features extracted using our method are clearly distinguishable in both VIS and NIR modalities.

V. Conclusion

In this paper, we present a novel forgery detection method based on spectrum spatial-temporal frequency clue. We then performed comprehensive experiments on three widely used VIS video forgery datasets and our newly constructed NIR video forgery dataset DeepfakeNIR. The experimental results show that our method outperforms existing methods on both ACC and AUC, and has excellent robustness and generalization ability in heterogeneous video forgery detection scenarios. In the future, we will try various backbone networks, not only the CNN model but also the transformer, and we will further explore the frame selection and block strategy to further expand and improve our method.

Near-Infrared (NIR) face forgery detection is an emerging research area that uses NIR imaging technology to detect manipulated or fabricated faces. Future research directions in NIR-based face detection include exploring invariant forgery clues between VIS and NIR modalities, developing few-shot learning methods for NIR-based deepfake detection, developing multi-modality approaches to face forgery detection etc. The proposed novel dataset DeepfakeNIR based on the NIR modality will further facilitate future research on forgery detection in the near-infrared modality.

REFERENCES

[1] Y. Zhu, Q. Li, J. Wang, C. Xu, and Z. Sun, “One shot face swapping on megapixels,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 4832–4842.
[2] S. Agarwal and H. Farid, “Detecting deep-fake videos from aural and oral dynamics,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2021, pp. 981–989.
[3] E. Burkov, I. Pasechnik, A. Grigorev, and V. Lempitsky, “Neural head reenactment with latent pose descriptors,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 13783–13792.
[4] L. Song, W. Wu, C. Fu, C. Qian, C. Change Loy, and R. He, “Pareidolia face reenactment,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 2236–2245.
[5] Z. Xu et al., “FaceController: Controllable attribute editing for face in the wild,” 2021, arXiv:2102.11464.
[6] C. Fu, Y. Hu, X. Wu, G. Wang, Q. Zhang, and R. He, “High-fidelity face manipulation with extreme poses and expressions,” IEEE Trans. Inf. Forensics Security, vol. 16, pp. 2218–2231, 2021.
[7] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, “Contrastive learning for unpaired image-to-image translation,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2020, pp. 319–345.
[8] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 8107–8116.

[9] H. Wang, H. Zhang, L. Yu, L. Wang, and X. Yang, “Facial feature embedded CycleGAN for VIS-NIR translation,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2020, pp. 1903–1907.

[10] Y. Wang, C. Peng, D. Liu, N. Wang, and X. Gao, “ForgeryNIR: Deep face forgery and detection in near-infrared scenario,” IEEE Trans. Inf. Forensics Security, vol. 17, pp. 500–515, 2022.

[11] N. Yu, L. Davis, and M. Fritz, “Attributing fake images to GANs: Learning and analyzing GAN fingerprints,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 7555–7565.

[12] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, “Not made for each other—audio-visual dissonance-based deepfake detection and localization,” in Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 439–447.

[13] B. Hosler et al., “Do deepfakes feel emotions? A semantic approach to detecting deepfakes via emotional inconsistencies,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2021, pp. 1013–1022.

[14] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, “Exploring temporal coherence for more general video face forgery detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 15024–15034.
[55] S. L. Happy, A. Dasgupta, A. George, and A. Routray, “A video database of human faces under near infra-red illumination for human computer interaction applications,” in Proc. 4th Int. Conf. Intell. Hum. Comput. Interact. (IHCI), Dec. 2012, pp. 1–4.

[56] 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.

[57] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1800–1807.

[58] H. H. Nguyen, J. Yamagishi, and I. Eichizen, “Capsule-forensics: Using capsule networks to detect forged images and videos,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 2307–2311.

[59] A. Khormali and J.-S. Yuan, “ADD: Attention-based DeepFake detection approach,” Big Data Cognit. Comput., vol. 5, no. 4, p. 49, Sep. 2021.

[60] C. Wang and W. Deng, “Representative forgery mining for fake face detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 14918–14927.

[61] H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, and N. Yu, “Multi-attentional deepfake detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 2185–2194.

[62] H. Lin, W. Luo, K. Wei, and M. Liu, “Improved Xception with dual attention mechanism and feature fusion for face forgery detection,” 2021, arXiv:2109.14136.

[63] D. Giera and E. J. Delp, “Deepfake video detection using recurrent neural networks,” in Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Survell. (AVSS), Nov. 2018, pp. 1–6.

[64] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, “A closer look at spatiotemporal convolutions for action recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6450–6459.

[65] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4724–4733.

[66] K. Hara, H. Kataoka, and Y. Satoh, “Learning spatio-temporal features with 3D residual networks for action recognition,” in Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW), Oct. 2017, pp. 3154–3160.

[67] A. Haliassos, K. Vougoukas, S. Petridis, and M. Pantic, “Lips don’t lie: A generalisable and robust approach to face forgery detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2021, pp. 5037–5047.

[68] Y. Luo, Y. Zhang, J. Yan, and W. Liu, “Generalizing face forgery detection with high-frequency features,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 16312–16321.

[69] S. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, “CNN-generated images are surprisingly easy to Spot...now,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 8692–8701.

[70] H. Wang, Z. Liu, and S. Wang, “Exploiting complementary dynamic incoherence for DeepFake video detection,” IEEE Trans. Circuits Syst. Video Technol., early access, Jan. 20, 2023, doi: 10.1109/TCVT.2023.3238517.

[71] 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. 2886–2895.

[72] 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.

Yukui Wang received the B.E. degree in computer science and technology from Hebei Normal University, Hebei, China, in 2020. He is currently pursuing the master’s degree with Xidian University. His current research interests include computer vision, pattern recognition, and machine learning.

Chunlei Peng (Member, IEEE) received the B.Sc. degree in electronic and information engineering from Xidian University, Xi’an, China, in 2012, and the Ph.D. degree in information and telecommunications engineering in 2017. From September 2016 to September 2017, he was a Visiting Ph.D. Student with Duke University, NC, USA. He is currently working with the School of Cyber Engineering, Xidian University. His current research interests include computer vision, pattern recognition, and machine learning.

Decheng Liu received the B.Sc. degree in electronic and information engineering from Xidian University, Xi’an, China, in 2016, and the Ph.D. degree in information and telecommunications engineering in 2021. He is currently working with the School of Cyber Engineering, Xidian University. His current research interests include computer vision and machine learning, especially for heterogeneous image analysis and its applications.

Nannan Wang (Member, IEEE) received the B.Sc. degree in information and computation science from the Xi’an University of Posts and Telecommunications in 2009 and the Ph.D. degree in information and telecommunications engineering in 2015. From September 2011 to September 2013, he was a Visiting Ph.D. Student with the University of Technology Sydney, NSW, Australia. He is currently working with the State Key Laboratory of Integrated Services Networks, Xidian University. He has published more than 100 articles in refereed journals and proceedings, including International Journal of Computer Vision, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON IMAGE PROCESSING, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. His current research interests include computer vision, pattern recognition, and machine learning.

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. Since 2001, he has been with the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor with the Ministry of Education, a Professor of pattern recognition and intelligent systems, and a Professor with the Chongqing University of Posts and Telecommunications, Chongqing, China. His current research interests include multimedia analysis, computer vision, pattern recognition, machine learning, and wireless communications. He has published five books and around 200 technical articles in refereed journals and proceedings. He is a fellow of the Institution of Engineering and Technology. He is on the editorial boards of several journals, including Signal Processing (Elsevier) and Neurocomputing (Elsevier). He served as the general chair/co-chair, the program committee chair/co-chair, and the PC member for around 30 major international conferences.