Hindawi
Security and Communication Networks
Volume 2023, Article ID 1648912, 10 pages
https://doi.org/10.1155/2023/1648912

Research Article

Local Corner and Motion Key Point Trajectory Extraction for Facial Forgery Identification

Qingtong Liu and Ziyu Xue

Academy of Broadcasting Science, NRTA, Beijing, China

Correspondence should be addressed to Qingtong Liu; qingtong_liu@126.com

Received 24 February 2022; Revised 2 August 2022; Accepted 11 April 2023; Published 15 May 2023

Academic Editor: Beijing Chen

At present, the development of deep forgery technology has brought new challenges to media content forensics, and the use of deep forgery identification methods to identify forged audio and video has become a significant focus of research and difficulty. Deep forgery technology and forensic technology play a mutual game and promote each other's development. This paper proposes a spatiotemporal local feature abstraction (STLFA) framework for facial forgery identification to solve the media industry challenges of deep forgery technology. To adequately utilize local facial features, we combine facial key points, key point movement, and facial corner points to detect forgery content. This paper establishes a spatiotemporal relation, which realizes face forgery detection by identifying abnormalities of facial keypoints and corner points for interframe judgments. Meanwhile, we utilize RNNs to predict the sequences from facial key point movement abnormalities and corner points for interframe. Experimental results show that our method achieves better performance than some existing methods and good anticompression forgery face detection performance on FF++.

1. Introduction

Media content forgery has brought some security problems to society. Especially with the development of autoencoders (AEs) [1] and generative adversarial networks (GANs) [2], media content forgery has become easy to achieve through deep forgery techniques. The techniques usually utilize deep learning methods to alter a person's identity in a video to synthesize a piece of media content that does not exist. Deep forgery identification techniques include both image-level detection and video-level detection.

Forgery detection of images or video frames is mostly the detection of forged video content, including color inconsistencies and semantic inconsistencies. Image forgery detection can be divided into detecting the image as a whole and detecting the facial area, according to the detection dimension. Forgery detection of the image as a whole is mainly to detect the physical properties of the image, such as the direction of the image's light source[3], the saturated pixel frequency [4], and the spectral sensitivity [4]. It is classified by judging the difference between forged and authentic images. Forgery detection for facial regions includes inconsistent iris color, missing tooth gaps, and inconsistent eye reflexes, including detection of facial artifacts using light estimation, global consistency and geometric estimation [5], corneal highlight region consistency detection [6], and facial artifact detection [7].

The detection of video sequences is mainly performed by combining optical flow anomalies, motion incoherence, or anomalies between video frames. Forgery detection based on optical flow mainly calculates the optical flow field of the target in the video and detects the inconsistency of the optical flow field [8]. Some authors utilize eye blinks [9], abnormal head movements [10], and facial distortions [11] to detect incoherent motion or abnormal behaviors in consecutive frames.

However, the early works were mainly focused on global features. Specifically, we notice that forgery detection features are particularly evident in key facial organs such as the eyes, nose, and mouth [5, 6, 12]. For example, Xue et al. [12]

found that only using facial organs such as the nose, lips, eyes, eyebrows, and chin can detect deep forgery very well.

Based on this, we first consider constructing the facial organs' relation. These organs can be abstracted to local features and represented by sequential vectors. We then adopt recurrent neural networks (RNNs) to capture their internal properties or differences to obtain instructive guidance that describes whether the face is falsified. For comprehensive detection, we realize face forgery detection for key facial local regions such as the lips, eyes, nose, eyebrows, and chin, thus achieving impressive performance. The contributions of our work are summarized as follows:

(1) We propose a spatiotemporal local feature abstraction (STLFA) framework for facial forgery identification, which establishes local features' relation via an organ-specific method.

(2) In STLFA, we combine abnormal facial movement detection and facial landmark time discontinuity detection to analyze the facial key point and corner point features frame by frame. Meanwhile, we judge video sequences' key point movement and corner point number transformation to achieve forgery identification of images and videos.

(3) This paper demonstrates the effectiveness and robustness of the proposed method and discusses and analyzes the advantages and disadvantages of STLFA.

2. Related Works

2.1. Deep Forgery Discrimination Based on Image or Video Frames. Currently, most forgery detection of images or video frames is performed by detecting manual features for forgery identification. The detection subject can be divided into two categories: image detection and inconsistency detection only for human faces.

Image forgery detection mainly detects the inconsistent lighting conditions and color inconsistencies in images. Chen et al. [13] proposed a robust dual-stream network by integrating dual-color spaces RGB and YCbCr using an improved Xception model, which considers both the luminance and chrominance components of dual-color spaces (RGB and YCbCr) to enhance the robustness. Johnson and Farid [3] proposed a method to detect lighting inconsistencies by estimating the direction of point light sources in a single image to estimate the consistency of light sources for the whole image. McCloskey and Albright [4] analyzed the structure of the popular GAN network. They found that the image generated by the GAN network differs from the captured image in color processing. They propose a method for forgery classification by saturated pixel frequency detection and spectral sensitivity detection.

The forgery detection of inconsistencies in the person's face focuses on the incomplete consideration of semantics in the content generation process by the deep forgery method, resulting in the generation of a person with inconsistent iris colors in the left and right eyes, inconsistent reflections, and uneven gaps in the teeth. Matern et al. [5] detected facial artifacts based on detecting intraframe image artifacts using light estimation, global consistency, and geometric estimation. Hu et al. [6] proposed a scheme to study whether the highlight patterns on the corneas of two eyes are consistent to determine whether they are fake. Li and Lyu [7] determined the forgery traces by detecting artifacts traced from the affine transformation during face forgery.

In order to integrate the features of facial regions, some authors proposed novel approaches. Wang et al. [14] proposed a method that fused facial region feature descriptor for forgery determination by extracting feature points of a person's face. Xue et al. [12] built a transformer model for a deepfake-detection method by organs to obtain the deepfake features. Yang et al. [15] proposed a method for detecting differences in face textures by amplifying the texture differences between genuine and fake images and using a bootstrap filter to enhance postprocessing-induced texture artifacts and display the underlying features of the artifacts.

2.2. Deep Forgery Discrimination Based on Video Sequences. The video sequence-based deep forgery approaches have more detection items than the image-based deep forgery approach. The forged video generation process is frame-by-frame leading to optical flow inconsistencies between the preceding and following frames and motion anomalies.

In terms of forgery identification based on optical flow detection, Amerini et al. [8] proposed a forgery detection method based on optical flow anomalies between different frames by extracting the correlation of the optical flow field and using a CNN classifier for classification. Trinh et al. [16] proposed a forgery detection framework by superimposing optical flow fields on RGB images for forgery detection. Caldelli et al. [17] proposed a CNN-based classification method to distinguish motion dissimilarities in the temporal structure of video sequences by using optical flow fields.

In terms of forgery identification based on abnormal motion detection, Li et al. [9] proposed a GAN-based model that could not represent blinking in fake synthetic videos, enabling the detection of blink inconsistencies. Yang et al. [10] proposed a detection method based on the inconsistency of 3D head pose estimation by extracting the coordinates of facial key points and calculating the direction vector difference between the center of the face and the coordinates of peripheral key points to achieve deep forgery detection. Sun et al. [11] proposed a geometric feature calibration module to determine the accuracy of interframe geometric features to determine the abnormal facial movements of characters.

3. Methods

3.1. Framework. In this section, we provide a detailed illustration of our proposed method. Figure 1 illustrates the architecture of STLFA. We used facial preprocessing modules to crop the eight facial organ regions, including the left eyebrow, right eyebrow, left eye, right eye, nose, mouth, inner mouth, and chin. We built a sequence group by facial

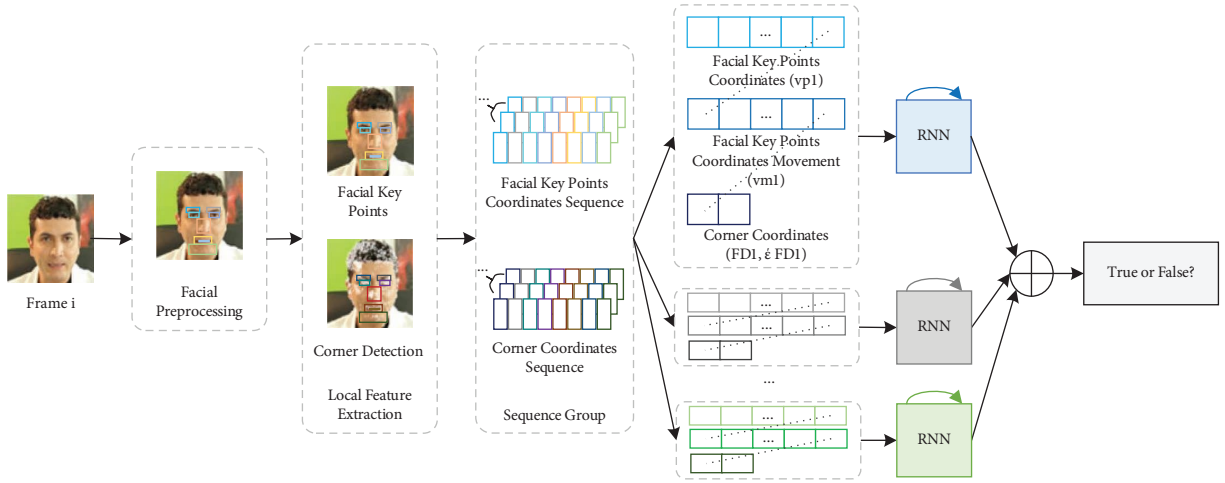


Figure 1: The framework of STLFA. The face image in this figure comes from the FF++ dataset [18] obtained from open access.

key points, key point movement, number of corner points, and number of variations. Meanwhile, RNN models are trained for each region until they have the detection ability. After that, we integrate the results from the RNNs and obtain the final prediction.

3.2. Facial Preprocessing. The facial preprocessing module mainly contains three steps: face detection, face landmark detection, and landmark alignment. Following [11], we use tracking and denoising methods to match the key points between video sequences to obtain the complete facial key point coordinates and coordinate movement. We utilized the Lucas–Kanade (LK) operation in the tracking method to track the coordinate points and forward-backward processes to eliminate inaccurate predictions. Meanwhile, the denoising method is used to solve the noise caused by the LK operation and to ensure the stability of the landmark, using the Kalman filter to integrate the prediction information.

3.3. Facial Key Points Extraction

3.3.1. Facial Key Points Coordinates Extraction. The facial key point coordinate detection method requires cropping the preprocessed image. After that, we detect 68 facial key points representing the facial shape, as shown in Figure 2(a). We select the key point frame to extract eight facial key organ regions based on the 68 key points, as demonstrated in Figure 2(b). We create vector v_p for each key organ region.

$$v_p = \left[v_{p_1}, v_{p_2}, \ldots, v_{p_8}\right]. \tag{1}$$

Each region can be expressed as v_{p_i}:

$$v_{p_i} = \left[x_i^1, y_i^1, x_i^2, y_i^2, \ldots, x_i^n, y_i^n\right], \tag{2}$$

where x_i^1 is the horizontal coordinate of the first key point in region i and y_i^1 is the vertical coordinate of region i.

3.3.2. Corner Extraction

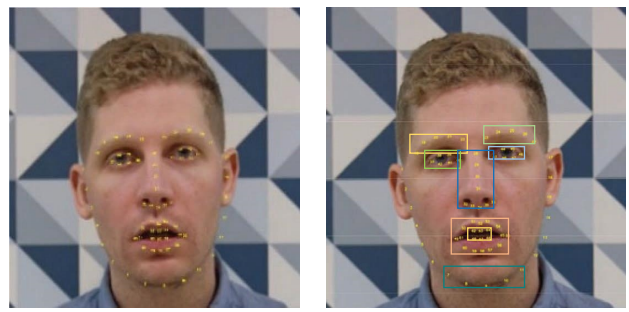
(1) Motivation for Using FAST Feature Points. The FAST algorithm is a corner detection algorithm mainly used to extract the feature points in the image. Based on the feature point information, the translation, distortion, and rotation objects in the dynamic process are associated with realizing the target tracking in a series of images of dynamic imaging and positioning. Wang et al. [14] found that although the fake video face was highly similar to the original video face, it still lost many fine details used to determine the FAST feature points and found that the phenomenon was more evident in the local area of the face. Based on this observation, we design a FAST feature descriptor to extract the phenomenon of the occasional failure of face-changing in the local area of the fake video and further complete the face forgery detection.

(2) Extraction Algorithm Feature Point of FAST. Features from accelerated segment test (FAST) [19] is an efficient corner point detection method mainly used for feature extraction of image corner points. The FAST method builds up the intensity of a pixel point I_p, sets the threshold value to t, and creates a Bresenham circle for 16-pixel points around p, as shown in Figure 3(a).

Designating pixel point p as a corner point if there is a set of n consecutive pixels in the circle that are all brighter than $I_p + t$ or darker than $I_p - t$.

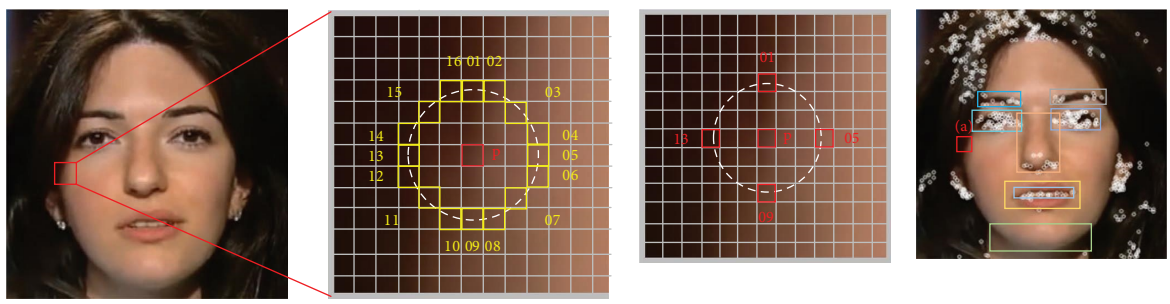
In order to speed up the operation, the pixel points compared with I_p can be simplified and set to 1, 5, 9, and 13, as shown in Figure 3(b). This paper focuses on establishing FAST corner point detection for eight regions extracted, such as the eyes, nose, lips, and eyebrows, and establishing corner point comparisons between frames, as shown in Figure 3(c).

We define pixel p as a corner when the circle in Figure 3(a) has a group of consecutive pixel points. Meanwhile,



(a) (b)

Figure 2: Facial key point coordinate detection: (a) 68 key points of facial contour. The face image is from FF++ [18]. (b) The key region cropped. The face image is from FF++ [18].



(a) (b) (c)

Figure 3: FAST feature point extraction algorithm: (a) p is the selected corner point, and a Bresenham circle is established around the point p. The face image is from FF++ [18]. (b) Simplified corner operations. The face image is from FF++ [18]. (c) Fast corner points detection in eight regions. The face image is from FF++ [18].

the points are brighter than $I_p + t$ or darker than $I_p - t$. In order to speed up the operation, the points can be simplified and only use points 1, 5, 9, and 13 to calculate, as shown in Figure 3(b). We focus on establishing FAST corner point detection for eight regions, as shown in Figure 3(c), and setting corner point comparisons between frames.

3.4. Abnormal Facial Movement Detection

3.4.1. Facial Shape Movement Abnormal Detection. Facial shape movement detection is based on the extraction of 68 feature points of facial feature extraction; the facial area is divided into 8 areas, and the temporal movement pattern of the feature points in each area is established for each area to realize facial shape movement abnormal detection. We analyze the movement of key points in each region and build a key points coordinate vector $v_{l_k}^i$.

$$v_{l_k}^i = \left[x_1^i, y_1^i, x_2^i, y_2^i, \ldots, x_{68}^i, y_{68}^i\right]. \tag{3}$$

The key point coordinate vector of the eight regions collection in frame i can be expressed as v_l^i:

$$v_l^i = \left[v_{l_1}^i, v_{l_2}^i, \ldots, v_{l_8}^i\right], \tag{4}$$

where $v_{l_1}^i \sim v_{l_8}^i$ represents the respective vectors of the eight regions in frame i and the corresponding key points are as follows: 6~10 represent the chin, 17~21 points represent the left eyebrow, 22~26 represent the right eyebrow, 36~41 represent the left eye, 42~47 represent the right eye, 27~35 represent the nose, 48~60 represent the mouth, and 61~67 stands for the inner mouth.

Then, we use $v_{l_1}^i \sim v_{l_8}^i$, extracted frame by frame, to provide clues for subsequent temporal discontinuity detection of facial motion morphology.

3.4.2. Facial Corner Abnormal Detection. Following [14], we use FAST to obtain feature points with a descriptor des of 32 dimensions. We assume that the number of corner points FD_i^k of the focus organ region k is in frame i, then FD_i^k can be expressed as follows:

$$FD_k^i = \sum_{k_i} \mathrm{des}[x], x \in [1, 32]. \tag{5}$$

In this way, a feature vector FD^i can be created for the eight regions:

$$FD^i = \left[FD_1^i, FD_2^i, \ldots, FD_8^i\right], \tag{6}$$

where FD_k^i is a statistical vector based on corner points in region i, containing the number of corner points in region k at frame i. We create time series based on FD_k^i to detect clues of alternating authentic and forgery faces in forgery videos.

3.5. Facial Landmark Time Discontinuity Detection

3.5.1. Facial Key Points Time Discontinuity Detection. We detect the temporal discontinuity of facial key point displacement between frames based on the displacement information of facial key points between consecutive frames. We analyze the movement of key points in each region and build a key point coordinate movement vector $v^i_{m_k}$; each region can be expressed as follows:

$$v^i_{m_k} = \left[\Delta x^1_i, \Delta y^1_i, \Delta x^2_i, \Delta y^2_i, \ldots, \Delta x^n_i, \Delta y^n_i\right]. \tag{7}$$

The key point coordinate movement vector of the eight regions collection in frame i can be expressed as v^i_m:

$$v^i_m = \left[v^i_{m_1}, v^i_{m_2}, \ldots, v^i_{m_8}\right], \tag{8}$$

where Δx^1_i is the adjacent frames variation in the horizontal coordinates; we can calculate Δx^1_i using $|x^1_{i+1} - x^1_i|$, the same as Δy^1_i.

3.5.2. FAST Feature Time Discontinuity Detection. The FD^i_k in Section 3.4.2 is the corner number vector of the described local region, and we use this vector to build the corner number difference vector ΔFD^k_i between consecutive frames:

$$\Delta FD^i_k = \left[FD^2_k - FD^1_k, FD^3_k - FD^2_k, \ldots, FD^i_k - FD^{i-1}_k\right]. \tag{9}$$

ΔFD^i_k is the difference between the number of corners in region k in frame i and the number in region k in frame $i-1$. The statistical vector ΔFD^i of the difference in the number of the corners in the whole facial region can be expressed as follows:

$$\Delta FD^i = \left[FD^i_1, FD^i_2, \ldots, FD^i_8\right]. \tag{10}$$

We use ΔFD^i to detect nonsmooth facial corner number changes in the video.

3.6. Facial Forgery Prediction

3.6.1. Facial Feature Vector Association. Based on $v^i_{l_k}$, $v^i_{m_k}$, FD^i_k, and ΔFD^i_k obtained in Sections 3.4 and 3.5, the local facial feature fusion vector $v^i_{f_k}$ is formed by concatenating the four types of feature vector sequences:

$$v^i_{f_k} = \left[v^i_{l_k}, v^i_{m_k}, FD^i_k, \Delta FD^i_k\right]. \tag{11}$$

Then, the local facial feature fusion vector for region k of the entire video can be expressed as follows:

$$v_{f_k} = \left[v^1_{f_k}, v^2_{f_k}, \ldots, v^n_{f_k}\right]. \tag{12}$$

We utilize a series of the local facial feature fusion vectors $v_{f_1} \sim v_{f_8}$ to represent the facial fusion features. After that, we use the connected feature vector v_{f_k} to train a dual-stream RNN model for each of the eight regions to classify the forgery videos.

3.6.2. RNN-Based Deep Forgery Detection. We utilize RNNs to model local facial feature sequences. In order to ensure an identical input dimension of the RNN and to achieve deep forgery detection at the video level, each video sample used as input is cut into a fixed length, and a fixed number of key frames are extracted. Based on the extraction results, the RNN parameters are selected for training to achieve deep forgery detection of the overall video.

Through the embedding process, the RNNs are adopted to model the feature sequences of each local region, learning the shape movement pattern, landmark difference pattern, and FAST feature point variation pattern. Then, the fully connected (FC) network is connected to each RNN output layer. Furthermore, calculate 8 FC layers output average result as the final prediction to achieve deep forgery detection based on the local regions of the face. We utilize F to represent this process:

$$F\left(R_1\left(v_{f_1}\right), R_2\left(v_{f_2}\right), \ldots, R_8\left(v_{f_8}\right)\right). \tag{13}$$

4. Experiments

4.1. Datasets

(1) FaceForensics++ (FF++) [18]: FF++ is one of the benchmark datasets for large-scale deep forgery detection, with a total of over 1,000 segments, more than 1.5 million frames in total, and over 1.5 TB of video data in the original video format. Meanwhile, a face detector is used to filter the video footage to ensure that there are three video qualities in the FF++ dataset, Raw, c23, and c40, characterized by many forged video segments, and a variety of deep forgery methods are considered.

(2) Celeb-DF [20]: The Celeb-DF (v2) dataset is a large-scale deepfake forensic dataset that addresses the shortcomings of poor forged video quality, apparent forgery traces, and flickering video faces. The Celeb-DF (v2) dataset improves the deep forgery generation method and the face key point localization method to obtain stable fake video content quality. The dataset contains 590 raw videos collected from YouTube with categories of different ages, races, and genders. 5639 HD deepfake videos are the same quality as the online broadcast videos.

(3) DFDC preview dataset [21]: This dataset comes from The Deepfake Detection Challenge hosted by Facebook. It is the preliminary dataset for the competition. It consists of 5,214 videos, of which the ratio of true and false content is 1 : 0.28, and forgery data contain data generated by two deep forgery methods. Each video is a clip of about 15 s.

4.2. Experiment Settings. During preprocessing, DLIB was used for face cropping and face landmark detection, and FAST detector and BRIEF descriptors were used for corner point detection and description. In the classification process, a bidirectional recurrent neural network connects to the feature sequences in the respective regions. Each RNN in the detection framework consists of a GRU (gated recurrent unit) with a hidden layer feature output dimension of 64. A dropout layer is set between the input and the RNN, using a fully connected network to connect to the output of the

RNN layer. Using two dr = 0.5 dropout layers separated between the RNN layer and the fully connected layer and inside, these experimental parameter settings partly refer to existing research results [22].

In the experimental dataset section, the ratio of training data to test data was 7 : 3, with 120 frames drawn from each video. The model was optimized using the Adam optimizer for the specific training process. We initialize the learning rate at 0.005, set the batch size to 1024, and the maximum number of iterations Epoch was 800 rounds. The experiments in this paper use AUC (area under curve) to evaluate the performance of the deep forgery detection model, and the AUC is calculated as follows:

$$\mathrm{AUC} = \frac{\sum \mathrm{pred}_{\mathrm{pos}} > \mathrm{pred}_{\mathrm{neg}}}{\mathrm{positiveNum} * \mathrm{negativeNum}}, \tag{14}$$

where $\mathrm{pred}_{\mathrm{pos}}$ is the predicted probability of getting a positive sample, $\mathrm{pred}_{\mathrm{neg}}$ is the predicted probability of getting a negative sample, positiveNum is the number of positive samples, negativeNum is the number of negative samples, and AUC is the number of samples where the predicted probability of a positive sample is greater than the predicted probability of a negative sample in the positiveNum ∗ negativeNum sample.

4.3. Experiments

4.3.1. Partial Organ Comparison. In this paper, experiments are conducted on the FF++ dataset to compare each organ region module's detection effect to verify each organ's region detection effect on deep forgery. In this paper, following the idea of [14], eight key regions such as the left eyebrow, right eyebrow, left eye, right eye, nose, mouth, inner mouth, and chin were set up and compared, as shown in Table 1. The "Points" results are obtained using facial key point coordinate detection and facial key point coordinate movement detection, "Coordinate" indicates the detection result using only the facial key point coordinates, and "Movement" indicates the detection result using only the facial key point movement coordinates. "C + M" indicates the result obtained by combining the key point coordinate detection and the facial key point coordinate movement detection. "Corners" is the result obtained using FAST corner number detection and corner number change detection. "All" means that the results of "Points" and "Corners" are combined with the experimental results of FAST features, and the RNNs of each segment are trained separately.

From Table 1, all local organs can be used individually in the FF++ dataset to detect whether the images contain forgeries. This paper observes that among the eight organ regions, the eyebrows, eyes, and mouth have the highest accuracy rate, while the nose and chin have a low accuracy rate. Also, in the "Points" detection group, where three experiments were set up, it was seen that "Coordinate" could perform a single-frame detection task with an average detection rate of 87.2%. "Movement" is the detection method combined with video sequences, with an average detection rate of 82.6%. The combination of "Coordinate" and "Movement" enables the combination of abnormal facial movement detection and facial landmark time discontinuity detection, allowing for more effective acquisition of key facial features with an accuracy rate of 91.1%.

4.3.2. Ablation Study. In this paper, we use the frame-level AUC to verify the effectiveness of face key point and corner point detection on deep forgery detection, respectively, to validate the proposed method. The models in the experiments are trained on FF++ (raw) and tested on three datasets: FF++, DFDC Preview, and Celeb-DF. The results are shown in Table 2.

The experimental results show that "Points" and "Corners" have similar detection results in terms of AUC, with an average of 71.3% and 74.1%, respectively, and all the best detection was achieved by "All," with an AUC of 75.9%. Meanwhile, in the FF++, DFDC Preview, and Celeb-DF datasets, the AUC values of "All" were higher than those of "Points" and "Corners" and "All" has a higher AUC than "Points" and "Corners." This proves that the method proposed in this paper, which combines facial key point and corner point detection, is reasonable and effective.

4.3.3. Comparison Experiments. In this paper, using frame-level AUC evaluation, we selected mainstream deep forgery detection methods based on full-frame face region forgery detection [18], fake face edge fusion region detection [23], facial landmark feature enhancement forgery and detection [11], visual distortion detection [24], and capsule network forgery detection [25]. Tests were carried out on datasets such as FF++, DFDC Preview, and Celeb-DF. We refer to the detection results of [11, 14], as shown in Table 3. In the FF++ dataset, "raw" represents the uncompressed data and "c40" represents the compressed LQ data.

As can be seen from Table 3, the AUC results of the proposed method on FF++ are better than those of mainstream methods such as Xception [18], Face X-ray [23], LRNet [11], DSP-FWA [24], and Capsule [25]. In particular, in the experimental group of "c40," the proposed method has better robustness for low-quality forged video identification, with a 1.7% improvement over LRNet [11] and a 35.8% improvement over Face X-ray [23].

In anticompression forgery face detection, our work shows a good forgery face detection performance. The method in this paper extracts the geometric features of the local facial region by combining the local facial key points and the corner. The extracted features have more robust and lower cost characteristics and have high sensitivity in detecting changes in the number of the corner. The strategy designed in this paper for face forgery detection through 8 local facial regions improves the accuracy of overall face forgery detection by reducing the detection error of a single region. The effectiveness of our strategy is also verified on FF++ (Raw, c40).

The low-complexity and high-performance geometric feature extraction method designed in this paper can effectively reduce the impact of image compression on the face

Table 1: Comparison table of local organs (Acc (%)).

Region attribute		Left eyebrow	Right eyebrow	Left eye	Right eye	Nose	Mouth	Inner mouth	Chin	Avg
Points	Coordinate	92.5	90.8	90.6	91.5	85.1	88.2	74.7	83.8	87.2
	Movement	83.4	82.7	84.5	83.2	80.2	84.3	82.4	79.8	82.6
	C + M	*93.4*	*91.3*	*91.6*	*92.7*	*87.2*	*90.1*	*94.7*	*87.8*	*91.1*
Corners		94.2	92.1	92.3	92.5	84.7	93.4	88.6	83.4	90.2
All		**97.2**	**97.7**	**98.8**	**98.3**	**96.4**	**98.6**	**95.1**	**94.3**	**97.0**

The bold values are used to highlight the results of the experiments conducted for this study. Specifically, they represent the performance of the proposed method in each experimental group.

Table 2: Ablation experiments (AUC (%)).

Datasets	FF++ (6284)	DFDC Preview (5214)	Celeb-DF (6819)	Avg
Points	99.2	56.2	58.4	71.8
Corners	96.7	62.7	63.1	74.5
All	**99.9**	**63.5**	**64.3**	**76.3**

The bold values are used to indicate the experimental records where the proposed method, discussed in this paper, demonstrated the most favorable outcomes within each experimental group. By highlighting these values in bold, we aim to emphasize the superior performance achieved by our method in those specific experimental conditions.

Table 3: AUC (%) results of the proposed method and mainstream methods on the FF++ dataset.

Methods	FF++	
	Raw	c40
Xception [18]	99.7	86.5
Face X-ray [23]	99.1	61.6
LRNet [11]	99.9	95.7
DSP-FWA [24]	93.0	—
Capsule [25]	96.6	—
Single XceptionNet [26]	—	**97.8**
Chen et al. [27]	99.92	95.2
SPSL [28]	—	82.8
PCL + I2G [29]	99.79	—
FTCN [30]	99.7	—
Lip forensics [31]	98.9	94.2
FDFL [32]	99.7	92.4
Ours	**99.9**	97.6

The bold values are used to highlight the experimental records that represent the most optimal performance within each experimental group. Specifically, the bold values labeled as "Ours" indicate the results obtained from the experiments conducted using our proposed method.

forgery detection task, and the experimental results further demonstrate this. We compared this method's training and testing results and other methods on the FF++ (Raw, c40) dataset in Table 3. The results show that our method achieves better performance than some existing methods, with a difference of 0.4% in AUC compared to the Single XceptionNet [26] method on FF++ (c40), and has better anticompression forgery face detection performance. The detection performance suffers less interference on c40 data.

4.3.4. Cross-Dataset Experiments. Our method can tolerate the local area detection, such as eyes, nose, and other organs, which is suitable for detecting the forgery videos with stain and shelter. To further demonstrate the robustness of our method, the models trained on FF++ (raw) were selected and tested on the DFDC Preview and Celeb-DF datasets. The results of training and testing on FF++ (raw, c40) in Table 4 sets cross-dataset experiments in individual organs and organ combinations.

The experimental results show that our method is innovative and can only use individual organs to detect forgery videos with defilement and stain. Meanwhile, using all organ regions has higher average accuracy. To further verify the ability of our method, we set up cross-dataset experiments to compare with the state-of-arts in Table 5.

The test results are shown in Table 5, Xception [18], LRNet [11], DSP-FWA [24], Capsule [25], Single XceptionNet [26], FWA [7], LipForensics [31], STIL [33], ADDNet-3D [34], and ours are compared. The method has certain advantages in the existing DFDC Preview cross-dataset test results, but the effect still needs to be further improved in the cross-dataset test results. The specific reasons are analyzed as follows: the framework of this paper utilizes the spatial and temporal features such as the spatial position of facial feature points and the statistical number of FAST corner points and shows good performance on the FF++ dataset. This paper strengthens the description and distinguishing capabilities of forgery faces to a certain extent by using geometric features and uses the RNN to model the

Table 4: The detection accuracy in cross-dataset experiments only uses local organs and organ combinations (Acc (%)).

Region attribute		FF++	CelebDF	DFDC Preview
Single attribute	Left eyebrow	97.2	63.9	61.2
	Right eyebrow	97.7	64.3	63.1
	Left eye	98.8	66.7	67.8
	Right eye	98.3	66.2	63.7
	Nose	96.4	61.8	60.3
	Mouth	98.6	66.4	64.1
	Inner mouth	95.1	59.3	57.9
	Chin	94.3	58.7	55.6
Multiattribute	Eyes	98.9	64.2	62.7
	Eyes + eyebrows	99.1	65.1	63.1
	Eyes + mouth	99.2	64.8	62.9
	Mouth + inner mouth	97.4	63.2	62.2
	Nose + mouth + inner mouth	97.8	63.9	62.4
	Mouth + inner mouth + chin	96.1	60.1	59.8
All		**99.4**	**65.8**	**63.7**

Table 5: Cross-dataset experiments (AUC (%)).

Methods	Celeb-DF	DFDC Preview
Train on FF++ (raw)		
Xception [18]	48.2	49.9
LRNet [11]	56.9	—
DSP-FWA [24]	64.6	—
Capsule [25]	57.5	53.3
FWA [7]	56.9	
Ours (raw)	**64.8**	**63.5**
Train on FF++ (c23)		
FWA [7]	53.9	
LipForensics [31]	**82.4**	—
Ours (c23)	65.1	**64.1**
Train on FF++ (c40)		
STIL [33]	**75.58**	—
ADDNet-3D [34]	60.85	
Ours (c40)	64.7	**63.8**

The bold values are used to highlight the experimental records that represent the most optimal performance within each experimental group. Specifically, the bold values labeled as "Ours" indicate the results obtained from the experiments conducted using our proposed method.

time series of features to complete fake face detection, which verifies the effectiveness of the framework. Applying geometric features improves the sensitivity to detecting facial feature point motion patterns and differential changes to a certain extent. Still, in the face of forging changes in the scene around the face of different datasets, the feature extraction method in this framework needs to be further optimized. Obtaining more effective forgery face features is the further optimization direction of this framework.

4.4. Discussion. Although the proposed method utilizes RNNs to model local facial feature sequences, it achieves deepfake discrimination through abnormal facial movement detection and facial landmark time discontinuity detection and exhibits good detection performance and compression resistance. Our method mainly mines the detection performance of each local face region for deep forgery and can effectively learn and model local face regions' forgery features and patterns. However, since the sample distribution of the FF++ dataset cannot represent all deep forgery techniques, the generalization of this method under the new data distribution is not explicitly guaranteed, which may lead to the degradation of performance in cross-database testing. Research on the generalization problem will be our future goal.

5. Conclusion

The development of deep forgery technology has brought new challenges to the authenticity of media content. The mutual promotion of deep forgery technology and forensics technology is prominent in addressing the challenges brought by deep forgery technology to the media industry. We focus on the consistency of facial key points and corner points' coordinates and propose a spatiotemporal local feature abstraction (STLFA) framework for facial forgery identification, which establishes local features' relation via an organ-specific method, which combines abnormal facial movement detection and facial landmark time discontinuity detection to analyze the facial key point, and corner point features frame by frame. It is mainly to detect the consistency of the movement of facial key point coordinates and the facial corner point number variations. At the same time, the method utilizes the bidirectional RNN to establish the sequence in eight local facial regions to model the facial shape pattern, the key point movement pattern, and the corner point number variations.

Experimental results show that our method performs better than some existing methods and achieves good anticompression forgery face detection performance on FF++. At the same time, for the detection of face forgery, the generalization ability under cross-dataset testing is also important. Therefore, a robust method with strong generalization ability is the goal of our future work.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the National Fiscal Expenditure Program of China under grant 130016000000200003.

References

[1] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013, https://arxiv.org/abs/1312.6114.

[2] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[3] M. K. Johnson and H. Farid, "Exposing digital forgeries by detecting inconsistencies in lighting," in *Proceedings of the 7th workshop on Multimedia and security*, pp. 1–10, New York, NY, USA, August 2005.

[4] S. McCloskey and M. Albright, "Detecting gan-generated imagery using color cues," 2018, https://arxiv.org/abs/1812.08247.

[5] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proceedings of the 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 83–92, IEEE, Waikoloa, HI, USA, January 2019.

[6] S. Hu, Y. Li, and S. Lyu, "Exposing GAN-generated faces using inconsistent corneal specular highlights," in *Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 2500–2504, IEEE, Toronto, Canada, June 2021.

[7] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," 2018, https://arxiv.org/abs/1811.00656.

[8] I. Amerini, L. Galteri, and R. Caldelli, "Deepfake video detection through optical flow based cnn," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, Seoul, Korea (South), October 2019.

[9] Y. Li, M. C. Chang, and S. Lyu, "In ictu oculi: exposing ai generated fake face videos by detecting eye blinking," 2018, https://arxiv.org/abs/1806.02877.

[10] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8261–8265, IEEE, Brighton, UK, May 2019.

[11] Z. Sun, Y. Han, and Z. Hua, "Improving the efficiency and robustness of deepfakes detection through precise geometric features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3609–3618, New Orleans, LA, USA, June 2021.

[12] Z. Xue, Q. Liu, H. Shi, R. Zou, and X. Jiang, "A transformer-based DeepFake-detection method for facial organs," *Electronics*, vol. 11, no. 24, p. 4143, 2022.

[13] B. Chen, X. Liu, Y. Zheng, G. Zhao, and Y. Q. Shi, "A robust GAN-generated face detection method based on dual-color spaces and an improved Xception," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3527–3538, 2022.

[14] G. Wang, Q. Jiang, and X. Jin, "FFR_FD: effective and Fast detection of DeepFakes based on feature point defects," 2021, https://arxiv.org/abs/2107.02016.

[15] J. Yang, S. Xiao, A. Li, G. Lan, and H. Wang, "Detecting fake images by identifying potential texture difference," *Future Generation Computer Systems*, vol. 125, pp. 127–135, 2021.

[16] L. Trinh, M. Tsang, and S. Rambhatla, "Interpretable and trustworthy deepfake detection via dynamic prototypes," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1973–1983, Waikoloa, Hawaii, USA, January 2021.

[17] R. Caldelli, L. Galteri, I. Amerini, and A. Del Bimbo, "Optical Flow based CNN for detection of unlearnt deepfake manipulations," *Pattern Recognition Letters*, vol. 146, pp. 31–37, 2021.

[18] A. Rossler, D. Cozzolino, and L. Verdoliva, "Faceforensics++: learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1–11, Seoul, Korea (South), October 2019.

[19] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," *European Conference on Computer Vision*, Springer, Berlin, Germany, 2006.

[20] Y. Li, X. Yang, and P. Sun, "Celeb-df: a large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3207–3216, Seattle, WA, USA, June 2020.

[21] B. Dolhansky, R. Howes, and B. Pflaum, "The deepfake detection challenge (dfdc) preview dataset," 2019, https://arxiv.org/abs/1910.08854.

[22] E. Sabir, J. Cheng, and A. Jaiswal, "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces*, vol. 3, no. 1, pp. 80–87, 2019.

[23] L. Li, J. Bao, and T. Zhang, "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5001–5010, Seattle, WA, USA, June 2020.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[25] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Use of a capsule network to detect fake images and videos," 2019, https://arxiv.org/abs/1910.12467.

[26] S. Cao, Q. Zou, X. Mao, Y. Dengpan, and W. Zhongyuan, "Metric learning for anti-compression facial forgery detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1929–1937, Virtual Event, China, October 2021.

[27] S. Chen, T. Yao, Y. Chen, S. Ding, J. Li, and R. Ji, "Local relation learning for face forgery detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, pp. 1081–1088, 2021.

[28] H. Liu, X. Li, and W. Zhou, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 772–781, Seattle, WA, USA, June 2021.

[29] T. Zhao, X. Xu, and M. Xu, "Learning self-consistency for deepfake detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15023–15033, Montreal, Canada, October 2021.

[30] Y. Zheng, T. Liu, Q. Li, and J. Li, "Integrated analysis of long non-coding RNAs (lncRNAs) and mRNA expression profiles

identifies lncRNA PRKG1-AS1 playing important roles in skeletal muscle aging," *Aging*, pp. 15044–15060, 2021.

[31] A. Haliassos, K. Vougioukas, and S. Petridis, "Lips don't lie: a generalisable and robust approach to face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5039–5049, Seattle, WA, USA, September 2021.

[32] J. Li, H. Xie, and J. Li, "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6458–6467, Seattle, WA, USA, June 2021.

[33] Z. Gu, Y. Chen, and T. Yao, "Spatiotemporal inconsistency learning for DeepFake video detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3473–3481, Stockholm, Sweden, September 2021.

[34] B. Zi, M. Chang, and J. Chen, "Wilddeepfake: a challenging real-world dataset for deepfake detection," in *Proceedings of the 28th ACM international conference on multimedia*, pp. 2382–2390, Seoul, Korea, October 2020.