electronics

MDPI

Article

LIDeepDet: Deepfake Detection via Image Decomposition and Advanced Lighting Information Analysis

Zhimao Lai [1,2], Jicheng Li [3,4], Chuntao Wang [5,*], Jianhua Wu [6] and Donghua Jiang [7]

[1] School of Immigration Administration (Guangzhou), China People's Police University, Guangzhou 510663, China; laizhimao@cppu.edu.cn
[2] School of Automation, Guangdong University and Technology, Guangzhou 510006, China
[3] School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China; 202010102003@mail.scut.edu.cn
[4] School of Forensic Science and Technology, Guangdong Police College, Guangzhou 510320, China
[5] College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, China
[6] School of Electronic Information Engineering, Jingdezhen Vocational University of Art, Jingdezhen 333000, China; jhwu@ncu.edu.cn
[7] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China; jiangdh8@mail2.sysu.edu.cn
* Correspondence: wangct@scau.edu.cn

Abstract: The proliferation of AI-generated content (AIGC) has empowered non-experts to create highly realistic Deepfake images and videos using user-friendly software, posing significant challenges to the legal system, particularly in criminal investigations, court proceedings, and accident analyses. The absence of reliable Deepfake verification methods threatens the integrity of legal processes. In response, researchers have explored deep forgery detection, proposing various forensic techniques. However, the swift evolution of deep forgery creation and the limited generalizability of current detection methods impede practical application. We introduce a new deep forgery detection method that utilizes image decomposition and lighting inconsistency. By exploiting inherent discrepancies in imaging environments between genuine and fabricated images, this method extracts robust lighting cues and mitigates disturbances from environmental factors, revealing deeper-level alterations. A crucial element is the lighting information feature extractor, designed according to color constancy principles, to identify inconsistencies in lighting conditions. To address lighting variations, we employ a face material feature extractor using Pattern of Local Gravitational Force (PLGF), which selectively processes image patterns with defined convolutional masks to isolate and focus on reflectance coefficients, rich in textural details essential for forgery detection. Utilizing the Lambertian lighting model, we generate lighting direction vectors across frames to provide temporal context for detection. This framework processes RGB images, face reflectance maps, lighting features, and lighting direction vectors as multi-channel inputs, applying a cross-attention mechanism at the feature level to enhance detection accuracy and adaptability. Experimental results show that our proposed method performs exceptionally well and is widely applicable across multiple datasets, underscoring its importance in advancing deep forgery detection.

Keywords: deepfake detection; imaging environment; intrinsic image decomposition; lighting inconsistency; cross-attention fusion

Citation: Lai, Z.; Li, J.; Wang, C.; Wu, J.; Jiang, D. LIDeepDet: Deepfake Detection via Image Decomposition and Advanced Lighting Information Analysis. *Electronics* **2024**, *13*, 4466. https://doi.org/10.3390/electronics13224466

Academic Editor: Cecilio Angulo

Received: 14 October 2024
Revised: 6 November 2024
Accepted: 11 November 2024
Published: 14 November 2024

1. Introduction

The rapid advancement of AI has facilitated highly realistic fake facial images and videos, known as Deepfakes. These deep learning-based image synthesis technologies, which swap identities or alter facial attributes, significantly challenge the societal trust traditionally based on the belief that "seeing is believing" [1]. Deepfakes are frequently exploited for nefarious purposes such as false news dissemination, online sexual content,

blackmail, and other illegal activities. Ensuring the authenticity of such evidence is therefore of utmost importance, underscoring the need for advanced verification technologies. When Deepfakes are maliciously introduced into legal proceedings, they can severely compromise the authenticity of evidence, thereby affecting the fairness of judicial processes and destabilizing social order. In response to these threats, researchers have focused on developing forensic methods for detecting and identifying forged images and videos [2–10]. However, the rapid evolution of Deepfake generation techniques and the limitations in detection methods' generalization pose significant challenges to practical application, despite these efforts.

Lighting conditions are crucial in determining the external imaging environment of videos, as they reflect the actual physical conditions during the imaging process. Despite advancements in computer graphics and generative models, accurately replicating real-world lighting remains a significant challenge. The imaging process of digital images, as depicted in Figure 1, involves several key components: the lens system, sampling filters, color filter arrays, imaging sensors, and digital image processors. Among these, the imaging sensor is the core component. When the camera shutter is activated, light from the natural scene traverses the lens, optical filters, and color filter arrays before reaching the imaging sensor. The sensor then converts the captured light into electrical signals via the photoelectric effect. An analog-to-digital (A/D) converter then digitizes these signals. The digital signals undergo further processing by the digital signal processor (DSP), which performs tasks such as white balance adjustment, image sharpening, gamma correction, and data compression. The final image is then stored in the camera's memory or displayed on the screen. In contrast, Figure 2 illustrates the process of image generation using GANs [11,12]. In this framework, the generative model creates samples based on a given random variable, while the discriminative model endeavors to differentiate between real and generated samples. Adversarial training continuously refines both models: the generative model aims to produce samples indistinguishable from real data, while the discriminative model increasingly struggles to differentiate between them.

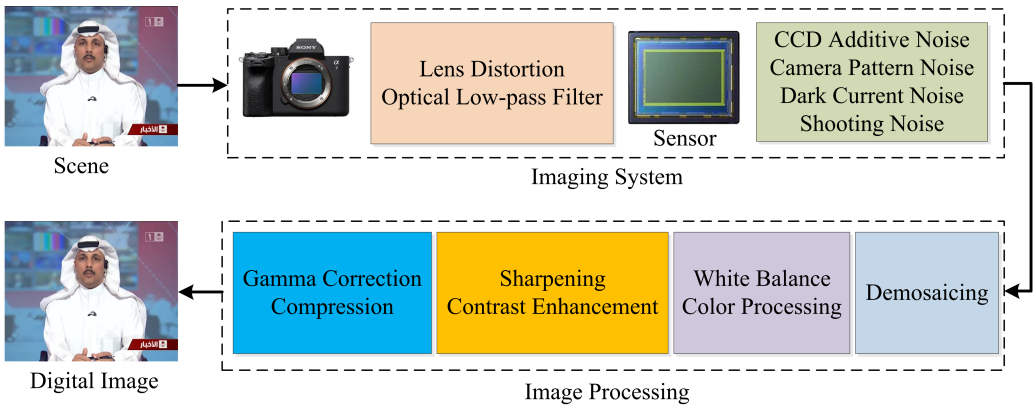

Figure 1. Imaging process of digital image.

The process outlined above underscores the significant differences in the external imaging environment between real and computer-generated images, which can be leveraged to identify distinguishing clues between genuine and fake images. Matern et al. [13] focused on detecting Deepfake videos by analyzing artifacts such as facial distortions and lighting changes during facial tracking and editing. The strength of this method lies in its minimal requirement for training data, as it targets common artifacts in the forgery process. However, it may encounter difficulties with meticulously crafted Deepfake videos that minimize these artifacts during generation. Li et al. [14] introduced a face-swapping video detection method based on lighting direction consistency. This approach calculates the two-dimensional lighting direction in videos and assesses the smoothness of lighting direction angle changes, offering low computational complexity and good real-time performance.

Nonetheless, it is sensitive to variations in lighting conditions, which can adversely affect detection accuracy. Gerstner et al. [15] described a real-time Deepfake detection technique that employs active illumination to induce controlled changes in the user's facial appearance. While this method facilitates real-time detection, it necessitates specific hardware and is vulnerable to video compression and post-processing effects. Wu et al. [16] developed a Deepfake detection algorithm based on block-wise lighting inconsistency. This method improves detection accuracy in forged videos by fusing channels to provide more lighting inconsistency information to the network's feature extraction layer, enabling block-wise similarity comparisons. However, it requires substantial computational resources and exhibits a higher false positive rate when processing non-forged videos. Zhu et al. [17] proposed a method to detect Deepfake videos by reconstructing and enhancing inter-frame inconsistencies, including lighting, color, and motion. This technique excels with complex video content but may lose accuracy when confronted with highly advanced forgery techniques. While these methods effectively utilize lighting inconsistency information for Deepfake detection, they often neglect the influence of environmental factors during filming. The lighting-related information that differentiates real from fake images can be significantly affected by varying shooting environments across different datasets, thereby impacting the generalization performance of detectors.

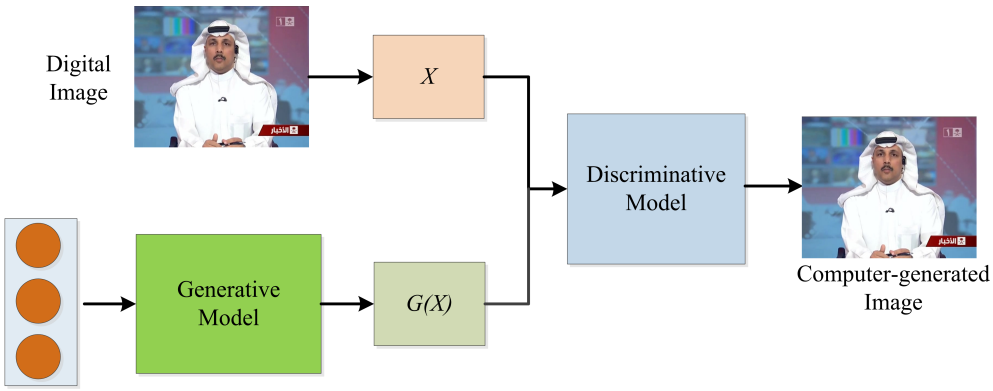

Figure 2. Process of image generation using generative adversarial networks.

In this paper, we explore the differences in lighting conditions between real and forged videos. Real videos typically maintain stable and consistent lighting throughout the recording process, whereas forged videos often rely on lighting models that generate less stable lighting across frames due to their limited simulation capabilities and frame-by-frame generation. We propose that by decomposing images and extracting features of lighting inconsistency, we can effectively differentiate between real and fake videos. Enhancing specific lighting details can highlight imaging differences between real and fake images, emphasizing inconsistencies. Conversely, other lighting information may introduce environmental interference, diminishing cross-domain detection performance, and should be suppressed to extract more intrinsic tampering traces. To address these challenges, we developed a lighting information extraction module based on color constancy principles to capture inconsistencies during imaging. To mitigate the impact of lighting changes, we incorporated a face material feature extraction module using the Pattern of Local Gravitational Force (PLGF) [18,19]. This module uses predefined convolutional masks to selectively process image patterns, filtering out illumination components to isolate face material features related to the reflectance coefficient. These texture-rich features effectively detect forgery. Additionally, we utilized the Lambertian lighting model to generate lighting direction vector maps for different frames, providing auxiliary information for temporal detection. Finally, we developed a multimodal learning framework combining CNN (EfficientNet B4) [20] and Vision Transformers (ViT) [21]. This framework integrates RGB images, face reflectance coefficient maps, lighting feature maps, and lighting direction

vector information as inputs from different modalities, employing a cross-attention fusion strategy at feature level to enhance detection generalization performance.

This paper's primary contributions are as follows:

- We propose a deep forgery detection method that utilizes image decomposition and lighting inconsistency to exploit the inherent differences in imaging environments between real and forged digital images, extracting reliable lighting information and effectively reducing interference from environmental factors to uncover more intrinsic tampering traces.
- We develop a hybrid multimodal learning method using Vision Transformers and CNN, treating image decomposition elements and lighting information as different modality inputs and employing a feature-level cross-attention fusion strategy to improve detection performance.
- Our experimental results show that our proposed algorithm consistently outperforms baseline algorithms across various datasets.

The structure of this paper is as follows: Section 2 reviews related works on Deepfake detection and intrinsic image decomposition. Section 3 outlines the proposed method, and Section 4 analyzes the experimental results. Finally, Section 5 concludes the paper.

2. Related Works

2.1. Deepfake Detection

Most Deepfake detection technologies today rely on deep learning frameworks, utilizing expert prior knowledge to aid neural networks in feature extraction. This includes analyzing inconsistencies in image blur levels, depth information, and light response non-uniformity. Given the significant differences between GAN-generated images and those processed by traditional methods, analyzing the unique structural traces left by GAN generators. For instance, differences in local textures and color components, as well as the checkerboard effect caused by deconvolution layers in GAN networks, serve as crucial indicators. These analyses form the foundation of Deepfake detection. Detection methods are categorized into two types: artifact-driven and vulnerability-driven. Artifact-driven methods focus on distinguishing real from fake by utilizing paired data for training, emphasizing artifacts in the forgery results. Early approaches employed handcrafted features, such as steganalysis features [22] or anomalies in blinking, head posture, pupil size, and dental details [13]. Deep learning methods soon became predominant. These can be further categorized based on the perspective of information extraction: (a) Spatial domain methods, like XceptionNet [23] and EfficientNet [20], extract feature information directly from the spatial domain of video frames. (b) Temporal domain methods, such as Two-branch [24], use a bidirectional LSTM [25] network to extract temporal information from consecutive frames and integrate it with other feature domains for detection. (c) Frequency domain methods, such as F3-Net [26], enhance artifacts from the Deepfake process through frequency-aware decomposition and local frequency statistics. The MPSM [27] method also merges spatial and frequency domain features to identify forgery traces. (d) Attention mechanism-based methods, such as Multi-Attention [28], approach video Deepfake detection as a fine-grained classification task and build upon the second-place method from the DFDC challenge. Additionally, methods analyzing biological signals, such as heart rate analysis [29], offer alternative perspectives for detecting Deepfake videos. Vulnerability-driven detection methods leverage specific traits of facial information carriers or inherent flaws in Deepfake generation processes. A representative method is Face X-ray [30] by Microsoft Research Asia, which detects fusion boundary traces required in the forgery process. Zhao et al. proposed Patch Consistency Learning (PCL) [31], predicting mask-like fusion operations, identifying faces with large masked areas as fake. Liu et al. [32] observed that multiple upsampling steps in the forgery process accumulate artifacts in the phase spectrum, leading to a phase spectrum-based detection method.

2.2. Intrinsic Image Decomposition

Image illumination decomposition is significantly applied in facial recognition, low-light image enhancement, and image relighting. Yang et al. [33] introduced a method using gray point estimation to determine light sources, enhancing color constancy. This approach is advantageous because gray points are common and easily detectable in natural images; however, it may underperform when gray points are scarce or indistinct. Fu et al. [34] introduced a weighted variational model to estimate reflectance and illumination simultaneously by optimizing a weighted energy equation, effectively balancing these estimations. Although this method performs well in complex scenes, it is computationally expensive. Hu et al. [35] introduced a fully convolutional network (Fc4) that improves color constancy using confidence-weighted pooling, effectively utilizing deep learning to handle large volumes of image data. Nonetheless, it may require extensive training data and might not be sufficiently sensitive to small objects or details. Hold-Geoffroy et al. [36] and Shi et al. [37] have each proposed deep learning-based methods for estimating outdoor lighting. These methods are notable for their ability to handle complex natural scenes and adapt to various lighting conditions. However, a significant limitation of these deep learning approaches is their reliance on a substantial amount of annotated data. Furthermore, the generalization capability of the models is constrained by the diversity of the training data. Guo et al. [38] proposed a method to enhance low-light images by estimating illumination maps. This approach simulates the human eye's adaptability to lighting changes to improve local details, although it may introduce noise in extremely dark areas. Baslamisli et al. [39] utilized convolutional neural networks (CNN) alongside reflectance and Retinex models for intrinsic image decomposition, successfully separating reflectance and illumination components, though this demanded substantial computational resources. Wang et al. [40] proposed a method using deep illumination estimation to enhance underexposed photos, improving image quality by estimating illumination components, yet facing challenges in high dynamic range scenes. Matern et al. [41] proposed a gradient-based method for detecting image forgery by analyzing gradient information to identify unnatural lighting changes. Guo et al. [42] introduced a zero-reference deep curve estimation method to enhance low-light image, which does not require additional reference image but may be less effective in extremely dark images. Ershov et al. [43] presented a physically feasible illumination distribution estimation method to improve white balance by predicting the scene's illumination distribution, offering a novel perspective for handling multi-light source scenes, though potentially necessitating more complex models for accurate prediction. Zhou et al. [44] introduced a low-light enhancement method based on the Retinex model, emphasizing structure preservation by utilizing the coefficient of variation (COV) to extract structural information. This approach achieved outstanding results in both subjective and objective evaluations.

3. Proposed Detection Methodologies

This section provides a detailed introduction to the proposed method, beginning with an overview of its architecture in Section 3.1. In Sections 3.2–3.5, we introduce the light information feature extraction module, the illumination normalization module, the light direction vector extraction module, and the feature-level cross-attention fusion strategy.

3.1. Overview

Several studies [14,16,17] have demonstrated that real videos maintain stable and consistent lighting components throughout the recording process, whereas forged videos exhibit lighting components generated by artificial lighting models. These models, due to their limited simulation capabilities and the frame-by-frame generation process, often result in unstable inter-frame lighting components in forged face videos. We propose a method to decompose images and extract lighting inconsistency features, enabling the distinction between real and fake videos. Our approach involves enhancing and amplifying lighting inconsistencies that reflect imaging differences between real and fake images. However,

some lighting information introduces environmental interference, which can reduce cross-domain detection performance. To mitigate this, we suppress and reduce such lighting information to extract more fundamental tampering traces. To effectively capture lighting inconsistency information during the imaging process, we design a lighting information feature extraction module based on the principle of color constancy. To minimize the impact of lighting changes, we incorporate a face material property feature extraction module based on PLGF. This module selectively processes different image patterns using predefined convolutional masks to filter out lighting components, resulting in face material property features related to the reflectance coefficient. These features, rich in texture information, serve as effective indicators for forgery detection. Furthermore, we employ the Lambertian lighting model to derive lighting direction vector maps across different frames, providing auxiliary information in the temporal domain. To enhance detection generalization performance, we design a multimodal learning method that integrates CNN (EfficientNet B4) and Vision Transformer (ViT). This method utilizes RGB images, face reflectance coefficient maps, lighting feature maps, and lighting direction vector information as inputs from different modalities. By employing a cross-attention fusion strategy at the feature level, our approach effectively combines these inputs. Figure 3 illustrates the architecture of our method.

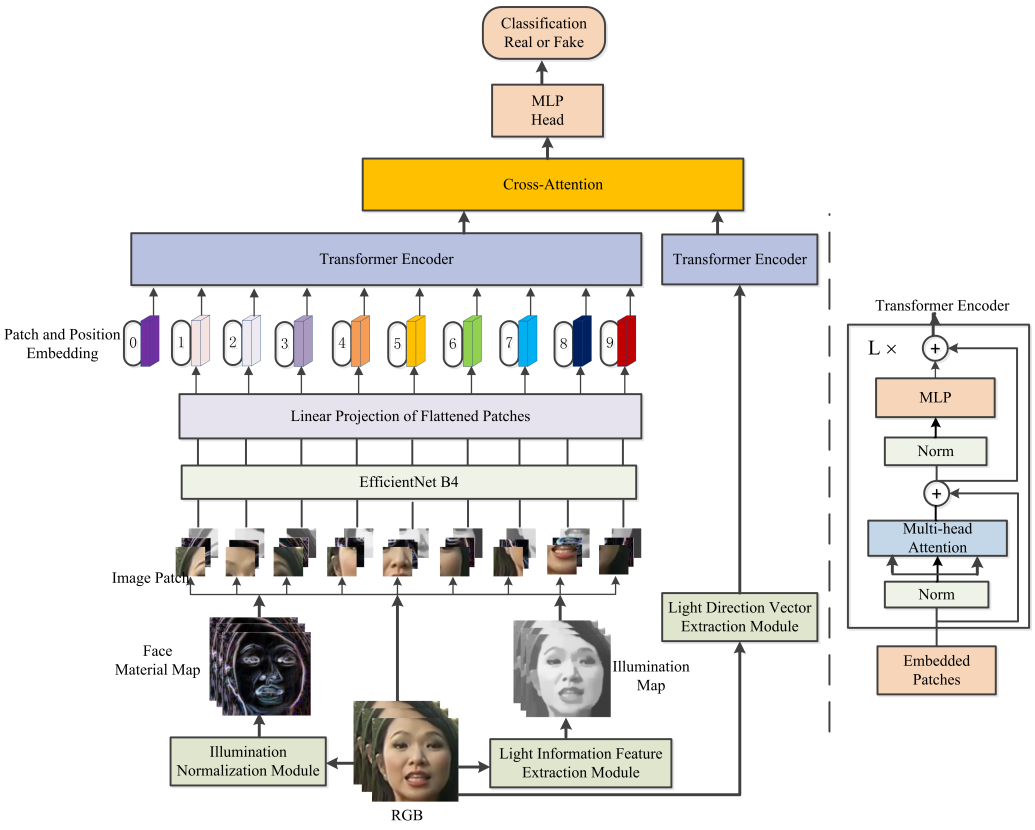

Figure 3. Architecture of the proposed method.

3.2. Light Information Feature Extraction Module

Color constancy is the human ability to maintain a consistent perception of an object's color despite variations in lighting conditions or external environments. In computer vision, color constancy algorithms are designed to simulate this human perceptual trait by removing the influence of lighting and external environments from the original image, thus extracting illumination information. Although Deepfake techniques can effectively swap the identity of a source image while retaining the target's attributes, they often alter the lighting conditions in the facial regions, leading to unnatural appearances. Abnormal lighting conditions can provide valuable supplementary information for detecting Deepfake.

The identity of the image is preserved while the lighting conditions in the facial region are exchanged, leading to unnatural effects.

According to the Retinex theory, an image $\mathbf{L}$ is composed of illumination image $\mathbf{M}$ and reflectance image $\mathbf{R}$:

$$\mathbf{L} = \mathbf{R} \circ \mathbf{M} \tag{1}$$

Initially, the illumination map $\hat{\mathbf{M}}$ is estimated by determining the maximum value among the R, G, and B color channels for each pixel:

$$\hat{\mathbf{M}}(x) \leftarrow \max_{c \in \{R,G,B\}} \mathbf{L}^c(x) \tag{2}$$

To preserve the image's overall structure and smooth texture details, the final illumination map is derived by solving the optimization problem in Equation (3):

$$\min_{\mathbf{M}} \|\hat{\mathbf{M}} - \mathbf{M}\|_F^2 + \alpha \|\mathbf{W} \circ \nabla \mathbf{M}\|_1 \tag{3}$$

In this context, α is a coefficient that balances the two terms. The symbols $\|\cdot\|_F$ and $\|\cdot\|_1$ denote the Frobenius norm and the ℓ_1 norm, respectively. Here, $\mathbf{W}$ represents a weight matrix, and $\nabla \mathbf{T}$ refers to a specific first-order derivative filter used in our experiments, which is the Sobel operator. The Sobel operator is a widely recognized discrete differential operator that computes an approximation of the gradient of an image intensity function. It is defined by the following 3×3 kernels for calculating the gradient in the x and y directions, respectively:

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad G_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{4}$$

These kernels are convolved with the input image to produce the gradient magnitudes in the horizontal and vertical directions, which are then combined to give a single gradient magnitude image.

To re-render faces with different attributes, the incident illumination must be transferred from the original image to the generated image. However, deep learning-generated faces often implicitly learn this model from the data, which can result in incorrect or imprecise estimates of the incident light, leading to artifacts. These artifacts typically manifest around the nose area, such as over-rendering on one side, as illustrated in the rightmost image of Figure 4.

Figure 4. Illustration of artifacts in deep learning-generated faces. The right-most image shows over-rendering around the nose area.

In deep learning, the generation of new faces through image interpolation involves the random selection of data points. Although the resulting images can often be described as qualitatively harmonious blends of different faces, they frequently exhibit a lack of global consistency. A notable example of this inconsistency is the variation in iris colors between the left and right eyes in many generated faces, as illustrated in Figure 5. This phenomenon, known as heterochromia, is quite rare in reality, thereby serving as a useful indicator for distinguishing real faces from generated ones.

Figure 5. Illustration of inconsistent iris colors in generated faces.

Based on Equations (1)–(3), we obtained the illumination information of the facial regions in real images and four forgery methods from the FF++ database [45] for the same frame. Figure 6 illustrates the visualization of these illumination maps.

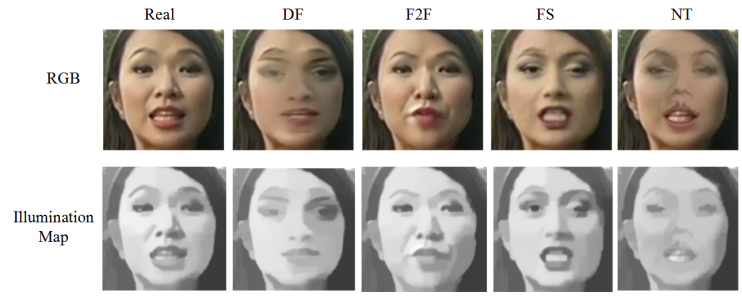

Figure 6. Visualization of illumination maps for real images and four forgery methods from the FF++ database.

3.3. Illumination Normalization Module

To mitigate the effects of illumination variations, we utilize a face material map extraction module based on the PLGF [18,19]. PLGF, originally developed for illumination-invariant face recognition, effectively filters out illumination components by selectively processing different image patterns using predefined convolutional masks. This process results in face material maps that are solely related to the reflectance coefficient, which contains rich texture information crucial for forgery detection. PLGF is an innovative dense local image descriptor comprising two feature components: local gravitational magnitude and local gravitational angle. Local gravitational magnitude quantifies irregular texture information, such as edges, lines, and corners, which correspond to the high-frequency part of the frequency domain. Consequently, the extracted face material maps serve as effective indicators for identifying forgeries. The local gravitational angle direction, on the other hand, represents the local geometric structure that is invariant to illumination (corresponding to the low-frequency part of the frequency domain). Most gradient-based descriptors currently use the difference between adjacent pixels (typically two pixels) to measure the direction and magnitude of the gradient. However, PLGF incorporates information from all surrounding pixels based on the direction and magnitude of the gravitational force. Therefore, this descriptor can capture more local information compared to other descriptors.

The PLGF descriptor is defined by Equations (4)–(6):

$$I_n = \arctan\left(\sqrt{\left(\frac{k_n * T_x}{k_n}\right)^2 + \left(\frac{k_n * T_y}{k_n}\right)^2}\right) \tag{5}$$

$$T_x(a,b) = \begin{cases} \frac{\cos(\arctan(b/a))}{a^2+b^2}, & \text{if } (a^2+b^2) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

$$T_y(a,b) = \begin{cases} \frac{\sin(\arctan(b/a))}{a^2+b^2}, & \text{if } (a^2+b^2) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

In these equations, k_n and I_n denote the original and processed images, respectively, with all mathematical operations in Equation (4) conducted on a per-pixel basis. The convolution operation, represented by $*$, is applied using two filter masks, T_x and T_y. Within these filter masks, the indices a and b indicate the relative positions with respect to the center of the mask. When the mask size is set to 5, the values for a and b range from -2 to 2.

By processing through the aforementioned illumination normalization module, we can obtain face material maps that are solely related to the reflectance coefficient. As shown in Figure 7, features rich in texture information effectively indicate forgery. The figure demonstrates that the eye and mouth regions of the face image exhibit more noticeable abnormal traces. This heightened visibility is attributed to the increased difficulty in forging these specific areas. Figure 8 visualizes face material maps for facial regions in real images and those altered by four forgery methods from the FF++ database [45], all for the same frame.

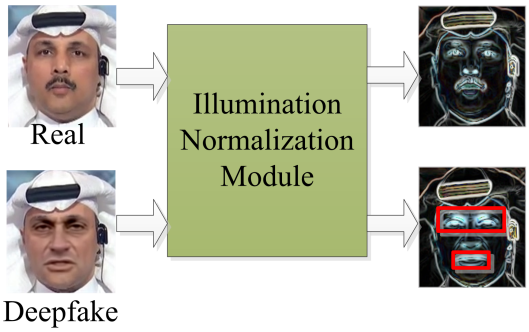

Figure 7. Face material map after illumination normalization. Abnormal traces in the eye and mouth regions are more noticeable.

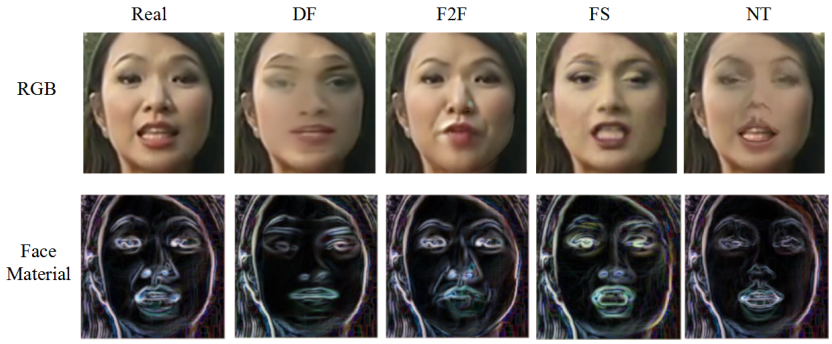

Figure 8. Visualization of face material maps for the facial regions in real images and four forgery methods from the FF++ database for the same frame.

3.4. Light Direction Vector Extraction Module

In this section, we apply the Lambert lighting model to estimate two-dimensional lighting direction vectors. The intensity of light reflected from an object's surface, I, is defined as follows:

$$I = k(\mathbf{N} \cdot \mathbf{L}) + \varepsilon \tag{8}$$

In this analysis, the diffuse reflectance of the object's surface is denoted by k, which is set to $k = 1$. The normal vector of the object's surface is represented by $\mathbf{N}$, while the lighting direction vector incident on the surface is denoted by $\mathbf{L}$. Additionally, ε signifies the ambient light intensity. Given that the actual lighting direction is a three-dimensional vector,

the theoretical intensity of light on the object's surface can be calculated using Equation (8). This theoretical intensity is crucial for understanding the interaction between light and the surface. The actual intensity of light, on the other hand, is represented by the grayscale pixel values of the object's surface. To accurately estimate the three-dimensional lighting direction vector **L**, one can minimize the difference between the theoretical and actual intensities of light. This approach ensures a precise alignment between the modeled and observed lighting conditions. The difference between the theoretical and actual intensities is described by the quadratic error function E. Figure 9 show the three-dimensional lighting direction vector, and Figure 10 show the two-dimensional lighting direction vector.

$$E(L,\varepsilon) = \left\| M \begin{pmatrix} L_x \\ L_y \\ L_z \\ \varepsilon \end{pmatrix} - \begin{pmatrix} I(x_1,y_1) \\ I(x_2,y_2) \\ \vdots \\ I(x_p,y_p) \end{pmatrix} \right\|^2 = \|\boldsymbol{M}\boldsymbol{v} - \boldsymbol{b}\|^2 \tag{9}$$

$$M = \begin{pmatrix} K_x(x_1,y_1) & K_y(x_1,y_1) & K_z(x_1,y_1) & 1 \\ K_x(x_2,y_2) & K_y(x_2,y_2) & K_z(x_2,y_2) & 1 \\ \vdots & \vdots & \vdots & \vdots \\ K_x(x_p,y_p) & K_y(x_p,y_p) & K_z(x_p,y_p) & 1 \end{pmatrix} \tag{10}$$

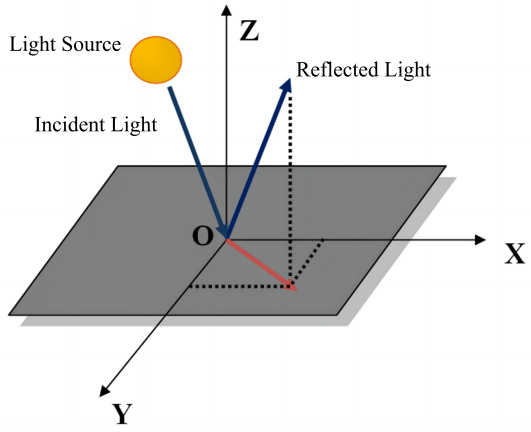

Figure 9. Three-dimensional lighting direction vector.

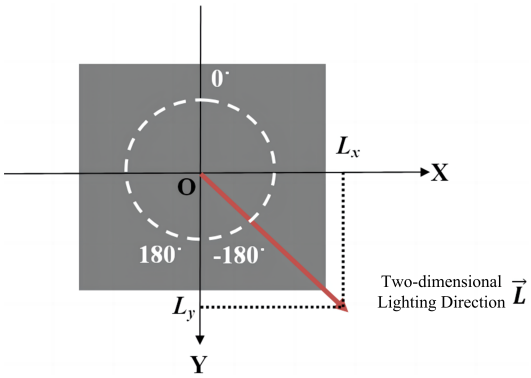

Figure 10. Two-dimensional lighting direction vector.

In this context, $\|\cdot\|$ represents the vector norm, which is a measure of vector magnitude. The three-dimensional lighting direction vector **L** is composed of components L_x, L_y, and L_z, corresponding to the lighting direction in the x, y, and z axes, respectively. The parameter p denotes the number of sampling points utilized to determine the lighting direction. At each i-th sampling point, the normal vector components are given by $K_x(x_i, y_i)$, $K_y(x_i, y_i)$, and $K_z(x_i, y_i)$, which represent the orientation in the x, y, and z directions, respectively.

Additionally, $I(x_i, y_i)$ signifies the actual intensity of light measured at the i-th sampling point. By minimizing the quadratic error function E in Equation (8), we can derive the three-dimensional lighting direction vector $\mathbf{L}_0$.

$$\mathbf{L}_0 = \begin{pmatrix} L_x \\ L_y \\ L_z \\ \varepsilon \end{pmatrix} = (\boldsymbol{M}^\top \boldsymbol{M})^{-1} \boldsymbol{M}^\top \boldsymbol{b} \tag{11}$$

In practical scenarios, the two-dimensional lighting direction is often calculated for images. In this case, we can set the z-component of $\mathbf{L}_0$ to zero to obtain the two-dimensional lighting direction vector $\mathbf{L}_1$, which is shown as follows:

$$\mathbf{L}_1 = \begin{pmatrix} L_x \\ L_y \\ \varepsilon \end{pmatrix} = (\boldsymbol{M}_1^\top \boldsymbol{M}_1)^{-1} \boldsymbol{M}_1^\top \boldsymbol{b} \tag{12}$$

$$\boldsymbol{M}_1 = \begin{pmatrix} K_x(x_1, y_1) & K_y(x_1, y_1) & 1 \\ K_x(x_2, y_2) & K_y(x_2, y_2) & 1 \\ \vdots & \vdots & \vdots \\ K_x(x_p, y_p) & K_y(x_p, y_p) & 1 \end{pmatrix} \tag{13}$$

Figure 11 shows the calculation process of the lighting direction. As shown in Figure 12, the intermediate results of the algorithm are demonstrated using frames 15, 20, 25, and 30 from the real video 1061402_A_001 in the DFDC database [46]. To assess the feasibility of estimating lighting direction vectors for detecting video forgeries, we conducted a preliminary statistical analysis using real and forged videos from existing public datasets. We assume that for short videos lasting several seconds, the lighting conditions in the shooting environment remain constant. Consequently, in naturally shot real videos, the lighting direction between adjacent frames tends to exhibit consistent patterns. In contrast, forged face-swapping videos are typically created by re-encoding sequences of independently altered video frames, which disrupts the natural alignment of lighting directions between adjacent frames. Therefore, significant variations in the lighting direction angles across a sequence of video frames may indicate potential forgery. When the calculation of lighting direction is confined to the facial area, such variations can specifically suggest a face swap operation. To validate the algorithm's effectiveness, we selected a real video consisting of 450 frames and its corresponding forged face-swapping version. We analyzed the distribution of their lighting direction angles, with the results depicted in Figure 13. As illustrated in Figure 13, the lighting direction angles in the real video change smoothly over time, whereas those in the forged video exhibit more abrupt fluctuations. This contrast supports the hypothesis that inconsistencies in lighting direction can serve as an indicator of video manipulation.

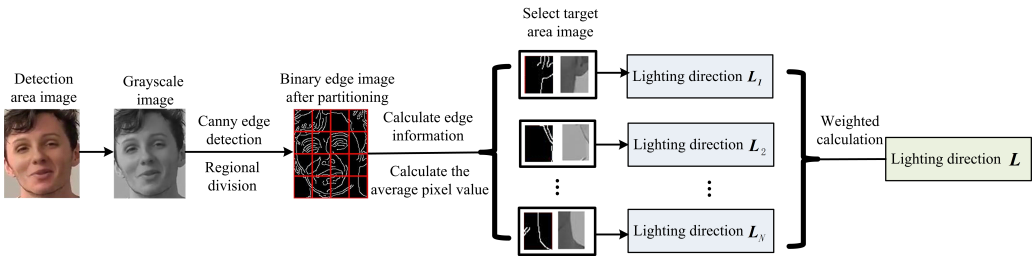

Figure 11. Calculation process of lighting direction.

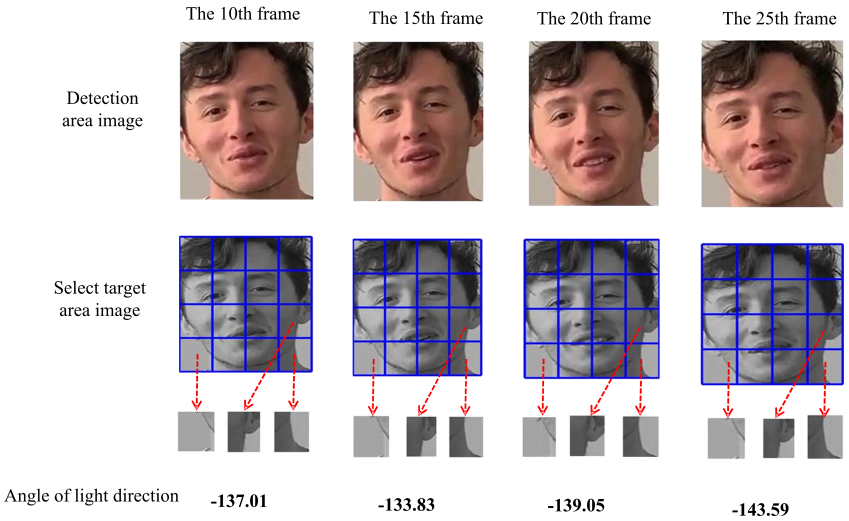

Figure 12. Calculation the angle of lighting direction.

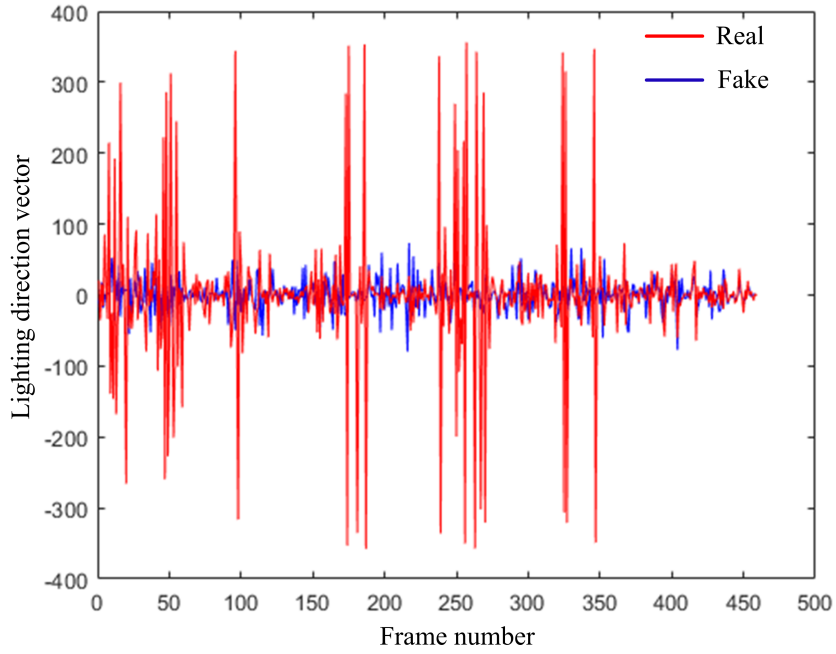

Figure 13. Comparison of lighting direction angles between real videos and their corresponding Deepfake videos.

3.5. Feature-Level Cross-Attention Fusion Strategy

ViT [21] is a model that leverages the Transformer [47] architecture, which is extensively employed for image classification tasks. The Transformer model's core module is structured around an encoder–decoder architecture, in which both the encoder and decoder are composed of multiple layers. The encoder's primary function is feature extraction, achieved through a combination of attention layers and fully connected layers. The attention mechanism within the encoder plays a crucial role by enabling the neural network to concentrate on pertinent information in the input data. It accomplishes this by assigning weights that determine the levels of attention, thereby minimizing the influence of irrelevant information. Subsequently, the decoder takes the extracted features and transforms them into the final output. Additionally, Transformer has the capability to model long-range dependencies and a broader receptive field, enabling it to more accurately extract and process image features. From a mathematical perspective, computing attention

involves mapping a query to a series of key–value pairs. This process primarily involves the following steps.

For the input sequence $\boldsymbol{z} \in \boldsymbol{R}^{N \times d}$, we project it using a linear transformation matrix $\boldsymbol{U}_{QKV}$ to obtain three vectors $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, representing the query vector, key vector, and value vector, respectively:

$$[\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}] = \boldsymbol{z}\boldsymbol{U}_{QKV}, \quad \boldsymbol{U}_{QKV} \in \boldsymbol{R}^{d \times 3d_h} \tag{14}$$

We compute the dot product of the query vector $\boldsymbol{Q}$ and the key vector $\boldsymbol{K}$, and scale it by a factor of $1/\sqrt{d_k}$ to obtain the similarity weight coefficients $f(\boldsymbol{Q}, \boldsymbol{K})$:

$$f(\boldsymbol{Q}, \boldsymbol{K}) = \frac{\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{K}}{\sqrt{d_k}} \tag{15}$$

We normalize the similarity weights using the softmax function and then compute the weighted sum of the value vectors to obtain the attention:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{16}$$

Multi-head attention involves applying h different linear transformations to the input $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, performing dot-product attention calculations, and then concatenating the results. The multi-head attention formula is given by the following:

$$\text{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Concat}(\text{head}_1, \cdots, \text{head}_h)\boldsymbol{W} \tag{17}$$

$$\text{where head}_i = \text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^{Q}, \boldsymbol{K}\boldsymbol{W}_i^{K}, \boldsymbol{V}\boldsymbol{W}_i^{V}) \tag{18}$$

In our multi-modal learning method based on Vision Transformer, we use RGB images, illumination maps, and face material maps as three modalities of input. To effectively fuse these multi-modal pieces of information and enhance the generalization performance of detection, we design an adaptive multi-modal adapter that employs a cross-attention fusion strategy at the feature level.

For each modality input, we first divide the image into N image patches (patches). These patches correspond to tokens in Natural Language Processing (NLP) and are arranged into a one-dimensional vector to serve as the input to the encoder. If the original input image has a resolution of $H \times W$ with C channels, and each image patch measures $n \times n$, then

$$N = \frac{H \times W}{n \times n} \tag{19}$$

Subsequently, positional encoding is applied to each image patch using sine or cosine curves. Positional encoding can be seen as a special embedding layer that encodes position information into a vector and adds it to the input data. This enables the model to differentiate information from various positions during attention computation, thereby enhancing its understanding of the input sequence's context.

To fuse information from different modalities, we design an adaptive multi-modal adapter. This adapter employs a cross-attention mechanism at the feature level, enabling features from different modalities to interact and enhance the model's representational capacity. The process involves the following steps:

Modality Feature Extraction: Each modality's input is processed through the ViT encoder to extract its feature representation.

Cross-Attention Fusion: Cross-attention mechanisms are used to fuse features from different modalities. Specifically, for the feature F_i of modality i, we compute its cross-attention with the feature F_j of modality j:

$$A_{ij} = \text{Attention}(F_i, F_j) \tag{20}$$

where Attention represents the attention computation function, which can be the standard self-attention mechanism.

Feature Fusion: The cross-attention mechanism fuses features from all modalities to produce the final multi-modal feature representation F:

$$F = \text{Concat}(A_{12}, A_{13}, A_{21}, A_{23}, A_{31}, A_{32}) \tag{21}$$

Finally, a fully connected layer outputs the classification result.

4. Experiments and Analysis

4.1. Datasets

To evaluate the generalization capability of the proposed method, we conducted experiments using several public benchmark datasets, including FaceForensics++ (FF++) [45], DeepFakeDetection (DFD) [45], Celeb-DF (CDF) [48], Deepfake Detection Challenge (DFDC) [46], and DeeperForensics1.0 (DF1.0) [49]. The FaceForensics++ dataset includes 1000 real videos sourced from YouTube, featuring frontal faces at resolutions of 480p or higher. It also contains 4000 Deepfake videos generated using Face2Face, FaceSwap, Deepfake, and NeuralTextures techniques, available in three compression levels: uncompressed (C0), compression rate 23 (C23), and compression rate 40 (C40). The DeepFakeDetection dataset consists of 363 high-definition videos recorded from 28 actors across 16 different scenarios, all with a resolution of 1080p. These videos serve as the basis for generating 3068 Deepfake videos. Unlike existing datasets, the DeepfakeDetection dataset not only ensures a wide range of actors but also introduces variations in several critical factors, including video focal length, background settings, subject movement, and facial expressions. The Deepfake Detection Challenge (DFDC) dataset presents a significant challenge due to its extensive scale and complexity. It comprises two versions: the preview dataset (DFDC-P) and the complete dataset (DFDC). This comprehensive approach ensures that the DFDC dataset closely mirrors the challenges encountered in practical scenarios, thereby providing a robust benchmark for evaluating Deepfake detection algorithms. Celeb-DF dataset is designed for Deepfake detection, containing 590 celebrity videos from YouTube and 5639 high-quality Deepfake videos. The dataset features varying ages, ethnicities, and genders, with face resolutions of 256 $\times$ 256 pixels. DeeperForensics-1.0 dataset comprises 100 actors from 26 countries, with a balanced distribution of ages and ethnicities and no facial obstructions. The video content includes various camera angles, facial expressions, and lighting conditions, all captured in 1080P resolution. In the synthesis process, real samples from the FF++ dataset are used as target videos, and various perturbations are applied to simulate real-world scenarios.

4.2. Evaluation Metrics

In this study, we utilize the Area Under the Receiver Operating Characteristic Curve (AUC) to evaluate the performance of our deepfake detection model. Deepfake detection is conceptualized as a binary classification problem, where samples are categorized as either positive (Deepfake) or negative (real). The effectiveness of binary classification models is commonly assessed using a confusion matrix, as demonstrated in Table 1. In this matrix, TP (True Positives) represents the number of positive samples correctly identified as positive, while FP (False Positives) denotes the number of negative samples incorrectly classified as positive. Conversely, FN (False Negatives) indicates the number of positive samples mistakenly classified as negative, and TN (True Negatives) refers to the number of negative samples accurately identified as negative. The detection accuracy of the model is derived from these components:

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \tag{22}$$

Table 1. Confusion matrix.

True Category	Predictive Category	
	Positive Sample	**Negative Sample**
Positive sample	TP	FN
Negative sample	FP	TN

TPR measures the proportion of actual positives correctly identified by the classifier, while FPR quantifies the proportion of actual negatives incorrectly classified as positives. These rates are calculated as follows:

$$FPR = \frac{FP}{FP + TN} \tag{23}$$

$$TPR = \frac{TP}{TP + FN} \tag{24}$$

AUC measures the likelihood that a classifier will rank a randomly selected positive sample higher than a randomly selected negative sample. A higher AUC value, approaching 1, signifies superior classification performance. One of the key advantages of AUC is its independence from the distribution of positive and negative samples, making it an objective measure of a model's classification effectiveness, particularly in binary classification tasks. Additionally, the calculation of AUC does not necessitate the setting of a threshold, further enhancing the objectivity of the evaluation.

4.3. Analysis of Results

Our experiment is conducted using the PyTorch 1.10.0 deep learning framework, executed on an NVIDIA GTX 1080 GPU platform. The training process involves a batch size of 50 and spans a total of 100 epochs. For training the model, we utilize the FF++ dataset. To evaluate the generalizability of our model, we perform cross-dataset evaluations using the DFDC, CDF, and DF1.0 datasets.

As demonstrated in Table 2, our method consistently outperforms previous approaches in most scenarios, including the overall average results. Notably, our method achieves substantial performance enhancements in terms of AUC, with improvements of 0.66% on the DFD dataset and 1.70% on the DF1.0 dataset, when compared to other methods like Xception [45], F3-Net [26], Face X-ray [30], SPSL [32], Two-Stream HF [50], LR-Net [51], and PCL [52]. Previous methods commonly suffer from dramatic performance drops when evaluated on DFDC and DF1.0, in which facial images are forged with more advanced manipulation techniques and then distorted with diverse postprocessing perturbations, while our method incorporates additional lighting inconsistency information into the neural network's feature extraction layer. This additional information helps the network better generalize across different types of face manipulations, thereby improving the overall performance and robustness of the detection algorithm.

In this section, we conduct a cross-dataset analysis using various subsets from the FF++ (C23) dataset. Our models are trained on samples from one of four distinct forgery techniques: DeepFakes (DF), FaceSwap (FS), Face2Face (F2F), and NeuralTextures (NT). Subsequently, we test the models on the remaining three techniques. The results, as shown in Table 3, reveal that our method outperforms other approaches in three out of the four scenarios. This finding underscores the effectiveness of our approach in distinguishing real videos from fake ones, regardless of the specific forgery techniques used. Consequently, this highlights the strong generalization capability of our method. The FF++ dataset may possess specific characteristics not adequately represented in other training datasets, leading to model overfitting on certain features and poor generalization to this dataset. The quality of dataset generation is inconsistent, making it challenging for models to detect high-quality fake videos that closely resemble real ones in visual cues. Additionally, the dataset may lack diversity in subjects, backgrounds, and lighting conditions, leaving

our model insufficiently trained to handle these variations. To address this, future work could incorporate data augmentation techniques to simulate more diverse manipulations, reducing overfitting and enhancing generalization. Alternatively, using models pre-trained on more diverse datasets and fine-tuning them on the Deepfake dataset could improve the model's ability to generalize to the specific features of this dataset.

Table 2. Frame-level (AUC(%)) on cross-datasets. The first and second rankings are shown in bold and underlined respectively.

Method	Train on FF++ (C23)				
	DFD	**DFDC**	**CDF**	**DF1.0**	***Cross Avg.***
Xception [45]	83.16	67.90	59.46	69.81	70.08
F3-Net [26]	90.09	73.26	65.17	70.78	75.58
Face X-ray [30]	85.60	70.01	74.20	72.60	75.50
SPSL [32]	83.23	75.56	76.88	71.41	76.77
Two-Stream HF [50]	<u>91.90</u>	**79.70**	79.40	73.80	<u>81.20</u>
LR-Net [51]	89.24	76.53	78.26	73.15	79.30
PCL [52]	90.05	73.23	81.80	<u>75.18</u>	80.82
MRL [53]	–	71.53	**83.58**	–	77.55
SFDG [54]	88.00	73.64	75.83	–	79.16
Ours	**92.56**	<u>78.36</u>	<u>80.23</u>	**76.88**	**82.00**

Table 3. Frame-level (AUC(%)) for cross-database evaluation across different manipulation types. The first rankings are shown in bold.

Training Set	Model	Testing Set (AUC)			
		DF	**F2F**	**FS**	**NT**
DF	Xception [45]	0.993	0.736	0.490	0.736
	Face X-ray [30]	0.987	0.633	0.600	0.698
	ReLAF-Net [6]	**0.997**	0.658	**0.721**	0.765
	Ours	0.990	**0.825**	0.550	**0.824**
F2F	Xception [45]	0.803	0.994	0.762	0.696
	Face X-ray [30]	0.630	0.984	0.938	0.945
	ReLAF-Net [6]	0.767	**0.995**	0.715	0.551
	Ours	**0.856**	0.994	**0.991**	**0.988**
FS	Xception [45]	0.664	0.888	0.994	0.713
	Face X-ray [30]	0.458	0.961	0.981	0.957
	ReLAF-Net [6]	0.783	0.775	0.988	0.621
	Ours	**0.725**	**0.994**	**0.996**	**0.983**
NT	Xception [45]	0.799	0.813	0.731	0.991
	Face X-ray [30]	0.705	0.917	0.910	0.925
	ReLAF-Net [6]	0.884	0.594	0.744	0.982
	Ours	**0.896**	**0.997**	**0.995**	**0.996**

4.4. Ablation Study

The ablation study presented in Table 4 systematically evaluates the contribution of each component within our LIDeepDet framework on the AUC scores across various datasets from the FF++(C23) training set. The results provide insights into the effectiveness of the Light Information, Illumination Normalization, Light Direction, and Cross-attention Fusion modules in detecting Deepfake. When only the Light Information module is enabled (first row), the model achieves an average AUC score of 71.20%, indicating that the light information feature extractor alone contributes significantly to the detection performance. The addition of the Illumination Normalization module (second row) improves the average AUC to 73.00%, demonstrating the benefits of using the PLGF-based feature extraction in enhancing the model's ability to distinguish between real and manipulated images. Further

enhancement is observed when the Light Direction module is included (third row), with an average AUC score of 76.90%. This suggests that the incorporation of lighting direction vectors, derived from the Lambertian lighting model, provides additional temporal context that aids in Deepfake detection. The most substantial performance boost is achieved when the Cross-attention Fusion module is integrated along with the other three modules (fourth row). The average AUC score soars to 82.00%, underscoring the importance of the cross-attention mechanism in fusing multimodal features and enhancing the model's generalization across different datasets. The cross-attention mechanism allows the model to focus on the most relevant features from each modality, leading to a more robust and accurate detection of Deepfake. In conclusion, the ablation study demonstrates that each component of the LIDeepDet framework plays a crucial role in improving the detection accuracy. The combined effect of these modules results in a significant enhancement in performance, highlighting the synergy between image decomposition and advanced lighting analysis in Deepfake detection.

Table 4. Comparison of AUC in ablation experiments (%). The first rankings are shown in bold. ✓ indicates the inclusion of the module.

Light Information	Illumination Normalization	Light Direction	Cross-Attention Fusion	Train on FF++(C23)				
				DFD	DFDC	CDF	DF1.0	*CrossAvg.*
✓				78.10	65.08	73.28	68.33	71.20
✓	✓			80.05	66.12	75.62	70.20	73.00
✓	✓	✓		85.56	72.33	77.25	72.45	76.90
✓	✓	✓	✓	**92.56**	**78.36**	**80.23**	**76.88**	**82.00**

4.5. Complexity Comparison

In this subsection, we compare the computational complexity of our proposed method, LIDeepDet, with the methods discussed in Section 2. Table 5 provides a summary of the complexity analysis.

Table 5. Complexity comparison of different methods.

Method	Time Complexity	Space Complexity	Real-Time Applicability
F3-Net [26]	$O(n^2)$	$O(n^2)$	Yes
Two-Stream HF [50]	$O(n \log n)$	$O(n)$	No
Ours	$O(n \log n)$	$O(n)$	Yes

We observe that LIDeepDet offers a competitive time complexity of $O(n \log n)$ and space complexity of $O(n)$, making it suitable for real-time applications. In contrast, Method A, which has a quadratic time complexity, is not suitable for real-time processing due to its high computational demands.

5. Conclusions

This paper presents LIDeepDet, an innovative Deepfake detection method that utilizes image decomposition and advanced lighting analysis to tackle the challenges posed by the rise of realistic AI-generated Deepfake images and videos. We employ a multi-faceted detection strategy that extracts robust lighting cues and mitigates environmental disturbances to uncover deeper-level alterations indicative of forgery. LIDeepDet marks a significant advancement in digital forensics by combining image decomposition with advanced lighting analysis. The method enhances detection accuracy by identifying inconsistencies in imaging environments and extracting reliable lighting information. By minimizing environmental interference, it enables precise identification of tampering traces, which is crucial for the reliability of digital evidence in legal settings. Integrating convolutional neural networks

(CNN) and Vision Transformers (ViT), it processes multi-channel inputs, including RGB images, face reflectance maps, and lighting features, leading to a more sophisticated detection capability. The cross-attention mechanism at the feature level enhances the model's generalization, ensuring adaptability across various datasets and scenarios. Experiments conducted on benchmark datasets such as FaceForensics++, DeepFakeDetection, Celeb-DF, DFDC, and DeeperForensics-1.0 demonstrate LIDeepDet's superior performance. The method outperforms existing algorithms, showing a notable increase in AUC scores, particularly on the DFD and DF1.0 datasets. Despite LIDeepDet's significant advancements, the rapid evolution of Deepfake generation techniques necessitates continuous research and development. Future efforts will focus on enhancing real-time performance, expanding applicability to other media manipulations, and integrating additional cues, such as biometric signals, to further strengthen detection capabilities.

Author Contributions: Data curation, Z.L. and J.L.; formal analysis, Z.L.; investigation, J.L.; methodology, Z.L. and J.L.; supervision, Z.L. and C.W.; visualization, Z.L. and C.W.; writing—original draft, Z.L. and C.W.; writing—review and editing, J.L., J.W. and D.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the National Fund Cultivation Project from China People's Police University under Grants JJPY202402, Scientific Research and Innovation Program for Young and Middle-aged Teachers from China People's Police University under Grants ZQN202411.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available on request to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Masood, M.; Nawaz, M.; Malik, K.M.; Javed, A.; Irtaza, A.; Malik, H. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Appl. Intell.* **2023**, *53*, 3974–4026. [CrossRef]
2. Xie, Y.; Cheng, H.; Wang, Y.; Ye, L. Domain Generalization via Aggregation and Separation for Audio Deepfake Detection. *IEEE Trans. Inf. Forensics Secur.* **2023**, *19*, 344–358.
3. Li, J.; Hu, Y.; Liu, B.; Gong, Z.; Kang, X. Boosting Deepfake Feature Extractors Using Unsupervised Domain Adaptation. *IEEE Signal Process. Lett.* **2024**, *31*, 2010–2014. [CrossRef]
4. Wang, H.; Liu, Z.; Wang, S. Exploiting complementary dynamic incoherence for deepfake video detection. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 4027–4040. [CrossRef]
5. Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; Li, H. Altfreezing for More General Video Face Forgery Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 4129–4138.
6. Gao, Y.; Zhang, Y.; Zeng, P.; Ma, Y. Refining Localized Attention Features with Multi-Scale Relationships for Enhanced Deepfake Detection in Spatial-Frequency Domain. *Electronics* **2024**, *13*, 1749. [CrossRef]
7. Lin, C.Y.; Lee, J.C.; Wang, S.J.; Chiang, C.S.; Chou, C.L. Video Detection Method Based on Temporal and Spatial Foundations for Accurate Verification of Authenticity. *Electronics* **2024**, *13*, 2132. [CrossRef]
8. Alhaji, H.S.; Celik, Y.; Goel, S. An Approach to Deepfake Video Detection Based on ACO-PSO Features and Deep Learning. *Electronics* **2024**, *13*, 2398. [CrossRef]
9. Gong, L.Y.; Li, X.J.; Chong, P.H.J. Swin-Fake: A Consistency Learning Transformer-Based Deepfake Video Detector. *Electronics* **2024**, *13*, 3045. [CrossRef]
10. Gao, Y.; Wang, X.; Zhang, Y.; Zeng, P.; Ma, Y. Temporal Feature Prediction in Audio–Visual Deepfake Detection. *Electronics* **2024**, *13*, 3433. [CrossRef]
11. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M. Generative Adversarial Networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
12. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
13. Matern, F.; Riess, C.; Stamminger, M. Exploiting visual artifacts to expose deepfakes and face manipulations. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 7–11 January 2019; pp. 83–92.

14. Li, J.; Liu, B.; Hu, Y.; Wang, Y.; Liao, G.; Liu, G. Deepfake video detection based on consistency of illumination direction. *J. Nanjing Univ. Aeronaut. Astronaut.* **2020**, *52*, 760–767.
15. Gerstner, C.R.; Farid, H. Detecting real-time deep-fake videos using active illumination. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 53–60.
16. Wenxuan, W.; Wenbo, Z.; Weiming, Z.; Nenghai, Y. Deepfake detection method based on patch-wise lighting inconsistency. *Chin. J. Netw. Inf. Secur.* **2023**, *9*, 167–176.
17. Zhu, C.; Zhang, B.; Yin, Q.; Yin, C.; Lu, W. Deepfake detection via inter-frame inconsistency recomposition and enhancement. *Pattern Recognit.* **2024**, *147*, 110077. [CrossRef]
18. Bhattacharjee, D.; Roy, H. Pattern of local gravitational force (PLGF): A novel local image descriptor. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 595–607. [CrossRef]
19. Li, Z.; Li, H.; Luo, X.; Hu, Y.; Lam, K.Y.; Kot, A.C. Asymmetric modality translation for face presentation attack detection. *IEEE Trans. Multimed.* **2021**, *25*, 62–76. [CrossRef]
20. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
21. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 30 April 2020.
22. Zhou, P.; Han, X.; Morariu, V.I.; Davis, L.S. Two-stream neural networks for tampered face detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1831–1839.
23. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
24. Masi, I.; Killekar, A.; Mascarenhas, R.M.; Gurudatt, S.P.; AbdAlmageed, W. Two-branch Recurrent Network for Isolating Deepfakes in Videos. In *Computer Vision—ECCV 2020, Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020, Proceedings, Part VII 16*; Springer International Publishing: Cham, Switzerland, 2020; pp. 667–684.
25. Jozefowicz, R.; Zaremba, W.; Sutskever, I. An empirical exploration of recurrent network architectures. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2342–2350.
26. Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; Shao, J. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Computer Vision—ECCV 2020*; Springer International Publishing: Cham, Switzerland, 2020; pp. 86–103.
27. Chen, S.; Yao, T.; Chen, Y.; Ding, S.; Li, J.; Ji, R. Local Relation Learning for Face Forgery Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, AAAI, Virtually, 2–9 February 2021; Volume 35, pp. 1081–1088.
28. Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; Yu, N. Multi-attentional Deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 2185–2194.
29. Ciftci, U.; Demir, I.; Yin, L. How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals. In Proceedings of the 2020 IEEE International Joint Conference on Biometrics (IJCB), Houston, TX, USA, 28 September–1 October 2020; pp. 1–10.
30. Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; Guo, B. Face X-ray for more general face forgery detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 5000–5009.
31. Zhao, T.; Xu, X.; Xu, M.; Ding, H.; Xiong, Y.; Xia, W. Learning to recognize patch-wise consistency for Deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 2185–2194.
32. Liu, H.; Li, X.; Zhou, W.; Chen, Y.; He, Y.; Xue, H.; Zhang, W.; Yu, N. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 772–781.
33. Yang, K.F.; Gao, S.B.; Li, Y.J. Efficient illuminant estimation for color constancy using grey pixels. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2254–2263.
34. Fu, X.; Zeng, D.; Huang, Y.; Zhang, X.P.; Ding, X. A weighted variational model for simultaneous reflectance and illumination estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2782–2790.
35. Hu, Y.; Wang, B.; Lin, S. Fc4: Fully convolutional color constancy with confidence-weighted pooling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4085–4094.
36. Hold-Geoffroy, Y.; Sunkavalli, K.; Hadap, S.; Gambaretto, E.; Lalonde, J.F. Deep outdoor illumination estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7312–7321.
37. Shi, W.; Loy, C.C.; Tang, X. Deep specialized network for illuminant estimation. In *Computer Vision—ECCV 2016, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016, Proceedings, Part IV*; Springer International Publishing: Cham, Switzerland, 2016; pp. 371–387.
38. Guo, X.; Li, Y.; Ling, H. LIME: Low-light image enhancement via illumination map estimation. *IEEE Trans. Image Process.* **2016**, *26*, 982–993. [CrossRef]

39. Baslamisli, A.S.; Le, H.A.; Gevers, T. CNN based learning using reflection and retinex models for intrinsic image decomposition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6674–6683.
40. Wang, R.; Zhang, Q.; Fu, C.W.; Shen, X.; Zheng, W.S.; Jia, J. Underexposed photo enhancement using deep illumination estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 6849–6857.
41. Matern, F.; Riess, C.; Stamminger, M. Gradient-based illumination description for image forgery detection. *IEEE Trans. Inf. Forensics Secur.* **2019**, *15*, 1303–1317. [CrossRef]
42. Guo, C.; Li, C.; Guo, J.; Loy, C.C.; Hou, J.; Kwong, S.; Cong, R. Zero-reference deep curve estimation for low-light image enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 1780–1789.
43. Ershov, E.; Tesalin, V.; Ermakov, I.; Brown, M.S. Physically-plausible illumination distribution estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 12928–12936.
44. Zhou, M.; Wu, X.; Wei, X.; Xiang, T.; Fang, B.; Kwong, S. Low-light enhancement method based on a Retinex model for structure preservation. *IEEE Trans. Multimed.* **2023**, *26*, 650–662. [CrossRef]
45. Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 1–11.
46. Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; Ferrer, C.C. The deepfake detection challenge (dfdc) preview dataset. *arXiv* **2019**, arXiv:1910.08854.
47. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 5998–6008.
48. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-df: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 3207–3216.
49. Jiang, L.; Li, R.; Wu, W.; Qian, C.; Loy, C.C. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 2889–2898.
50. Wu, J.; Zhang, B.; Li, Z.; Pang, G.; Teng, Z.; Fan, J. Interactive two-stream network across modalities for deepfake detection. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 6418–6430. [CrossRef]
51. Sun, Z.; Han, Y.; Hua, Z.; Ruan, N.; Jia, W. 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, Nashville, TN, USA, 19–25 June 2021; pp. 3609–3618.
52. Zhao, T.; Xu, X.; Xu, M.; Ding, H.; Xiong, Y.; Xia, W. Learning self-consistency for deepfake detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 15023–15033.
53. Yang, Z.; Liang, J.; Xu, Y.; Zhang, X.; He, R. Masked Relation Learning for Deepfake Detection. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 1696–1708. [CrossRef]
54. Wang, Y.; Yu, K.; Chen, C.; Hu, X.; Peng, S. Dynamic Graph Learning with Content-Guided Spatial-Frequency Relation Reasoning for Deepfake Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 7278–7287.