Expert Systems With Applications 297 (2026) 129475

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Dual-branch hierarchical feature fusion network for video source camera identification

Bo Wang [a], Jiaqi Chi [a], Zhuocheng Wu [b], Huimin Liu [a], Wei Wang [c,*]

[a] School of Information and Communication Engineering, Dalian University of Technology, No.2 Linggong Road, Ganjingzi District, Dalian, 116024, Liaoning, China
[b] School of Computer Science and Technology, Dalian University of Technology, No.2 Linggong Road, Ganjingzi District, Dalian, 116024, Liaoning, China
[c] Institute of Automation, Chinese Academy of Sciences, No. 95 Zhongguancun East Road, Haidian District, Beijing, 100190, Beijing, China

ARTICLE INFO

Keywords:
Source camera identification
Video forensics
Transformer block
Convolutional Neural Network
Feature fusion

ABSTRACT

With the rapid development of video platforms, digital video has become one of the core mediums for information dissemination. However, the extensive circulation of video content has also given rise to various social issues, including extortion, fraud, and misinformation. In this context, Video Source Camera Identification (VSCI), as a key technology in video forensics, plays an irreplaceable role in combating the spread of false information and assisting in crime identification. Traditional Source Camera Identification (SCI) methods primarily rely on various trace features generated during the capture process. However, with the popularization and advancement of image processing applications, extracting these feature traces has become increasingly challenging. Furthermore, due to storage and transmission limitations, video data often requires multiple compression processes. This multiple compression not only destroys the original features of the video but also introduces complex noise interference, making video source identification significantly more challenging than SCI for images. While Convolutional Neural Networks (CNNs) excel at local feature extraction, their ability to capture global information is limited, which constrains their identification performance in complex scenarios. We propose a dual-branch hierarchical feature fusion network structure to address the issue. The network extracts local and global features using CNN and Transformer, respectively, and achieves efficient feature fusion through a hierarchical feature fusion module, thereby comprehensively enhancing identification performance. To verify the feasibility and effectiveness of the proposed method, we conducted experiments using VISION and QUFVD dataset. The experimental results demonstrate excellent identification performance of this method.

1. Introduction

With the rise of video platforms and the widespread adoption of smartphones, users can share content online ubiquitously. However, this convenience has introduced societal issues like extortion, fraud, and misinformation. Digital images/videos have become increasingly important as crucial evidence and historical records, making their authenticity verification a core issue requiring urgent attention. Source Camera Identification (SCI) technology, as a crucial means for verifying digital content authenticity, can precisely match videos with recording devices and track device models involved in cases, thereby effectively curbing illegal activities such as unauthorized recording and copyright infringement, and maintaining social order and public safety.

Although numerous methods have achieved significant results in source camera identification for image (Berdich et al., 2023; Rana et al., 2023; Sychandran & Shreelekshmi, 2024; Wang et al., 2024c), with the rapid rise of short-video platforms, the demand for forensic analysis of illegal videos are growing increasing. Although there is some overlap between the video generation process and the image generation process, their technical challenges are more severe. Videos are essentially sequential data composed of continuous frame images that must undergo compression processing due to storage and transmission limitations. Moreover, internet platforms typically perform secondary compression on uploaded videos, and this multiple compression process makes the noise interference in videos more complex, resulting in video source identification being significantly more challenging than image source identification, as shown in Fig. 1. Therefore, developing efficient multimedia forensic technologies to address various video source identification needs has become crucial in solving illegal video-related issues.

In recent years, Convolutional Neural Networks (CNNs) have made significant progress in the task of Video Source Camera Identification

* Corresponding author.
E-mail addresses: bowang@dlut.edu.cn (B. Wang), 0703chi@mail.dlut.edu.cn (J. Chi), 15668854383@163.com (Z. Wu), 1292202003@qq.com (H. Liu), wwang@nlpr.ia.ac.cn (W. Wang).

https://doi.org/10.1016/j.eswa.2025.129475
Received 30 April 2025; Received in revised form 13 August 2025; Accepted 20 August 2025
Available online 23 August 2025

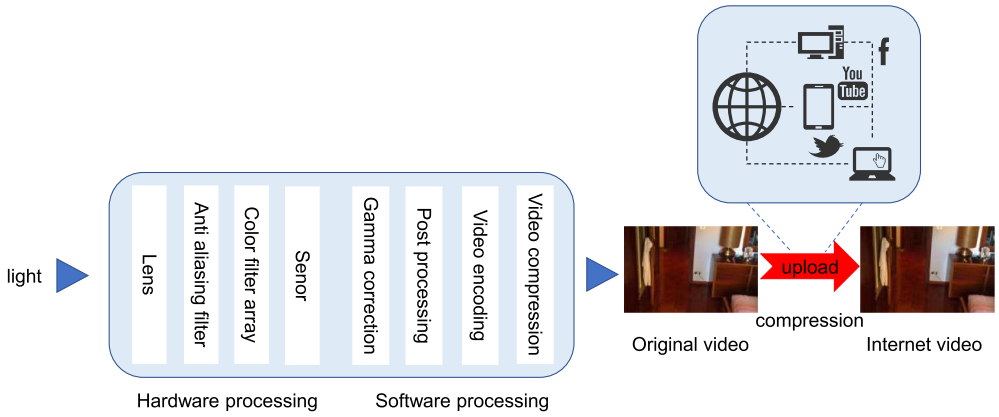

Fig. 1. Digital content generation process within and beyond digital cameras (Bennabhaktula et al., 2022).

(VSCI) (Anmol & Sitara, 2024; Bennabhaktula et al., 2022; Manisha et al., 2023). Although CNNs perform well in local feature extraction, their ability to capture global context is limited. Meanwhile, the Transformer demonstrated unique advantages in modeling long-range dependencies with its self-attention mechanism (Vaswani et al., 2017), and showed great potential in the field of VSCI (Elharrouss et al., 2025; Lu et al., 2024).

However, the existing methods often face two key limitations: (1) The traditional dual-branch architecture usually adopts simple feature concatenation or addition methods to fuse global and local features, failing to fully consider the hierarchical correlation of multi-scale features; (2) Transformer and CNNs have inherent limitations in local/global feature extraction respectively, which restricts their further application in SCI tasks. To address these issues, we propose a Dual-branch Hierarchical Feature Fusion Network (DHFFNet) structure. This network architecture innovatively extracts global and local features respectively using Swin-T network and Multi-Stage CNN, and achieves adaptive fusion of global and local features through the Feature Fusion Module (FFM). The trained model is then used to classify the frames, predicting the source device of each frame, and compiling the frames belonging to a single video through majority voting to predict the source camera device of each video. The experimental results show that this method has achieved superior performance compared to the current SOTA methods on datasets such as VISION and QUFVD, providing a new technical approach for VSCI.

In this paper, we introduce the Dual-branch Hierarchical Feature Fusion Network (DHFFNet) method to address the challenges of VSCI. The main contributions of this paper can be summarized as follows:

- We propose a dual-branch hierarchical feature fusion network that effectively captures both local and global features at different scales. By leveraging the complementary strengths of local and global feature extraction, it addresses the limitations of traditional methods.
- The model integrates Gated Spatial Attention Unit (GSAU), Channel Attention (CA), Spatial Attention (SA), and Local Importance-based Attention (LIA), achieving adaptive fusion of global and local features. This attention mechanism ensures that the model can dynamically focus on the most relevant features for identification.
- The proposed DHFFNet model has achieved promising results on both VISION and QUFVD datasets, demonstrating its identification capabilities.

The paper is organized as follows. Section 2 provides a review of existing methods for VSCI. Section 3 introduces the proposed network architecture. In Section 4, we demonstrate the nice performance of the proposed method for VSCI using datasets. Finally, we conclude the paper with Section 5, with acronyms listed in Table 1.

Table 1
Table of expansion and their abbreviations.

Expansion	Abbreviation
Video Source Camera Identification	VSCI
Electric Network Frequency	ENF
Convolutional Neural Networks	CNNs
Source Camera Identification	SCI
Swin Transformer	Swin-T
Gated Spatial Attention Unit	GSAU
Channel Attention	CA
Spatial Attention	SA
Local Importance-based Attention	LIA
Photo Response Non-Uniformity	PRNU
Inverted Residual Multi-layer Perceptron	IRMLP
Charge Coupled Device	CCD
Window Multi-head Self-Attention	W-MSA

2. Related work

In this section, we will discuss various methods for video source camera identification and their advantages and disadvantages. Over the years, researchers and digital forensics experts have invested considerable time in designing VSCI methods from different perspectives. We will analyze in depth the strengths and limitations of three categories of VSCI methods, including methods based on metadata, Photo Response Non-Uniformity (PRNU) and deep learning.

2.1. Identification methods based on metadata

Metadata-based methods perform SCI by analyzing metadata information. For instance, Ngharamike et al. (2023) proposed a technique leveraging Electric Network Frequency (ENF) to perform VSCI containing ENF. Similarly, Yang et al. (2020) utilized video container structures to trace video sources. Although metadata-based approaches have shown notable success, the widespread adoption of video editing software has made metadata manipulation increasingly accessible, significantly compromising their performances in Video Source Camera Identification (VSCI) (Li et al., 2024).

2.2. Identification methods based on PRNU

Early research primarily combined camera imaging characteristics with traditional algorithms, focusing on the inherent properties of the cameras themselves. By analyzing and modeling these characteristics, researchers could perform source camera identification by analyzing and comparing their feature values. A series of source camera identification methods were derived from these characteristics. Kurosawa et al. (1999) discovered Photo Response Non-Uniformity (PRNU) in images,

which is a unique characteristic inherent to cameras. The presence of PRNU originates from dark current non-uniformity in Charge-Coupled Device (CCD) sensors, generating fixed pattern noise. As a unique camera fingerprint, PRNU ensures higher identification accuracy and reliability, thereby providing robust evidence for SCI.

Yang et al. (2021) proposed a PRNU based on the video forensic method that considers video I-frames to improve processing time and accuracy. Lawgaly et al. (2022) proposed a weighted average-based video PRNU estimation method that extracts noise residuals from each video. The estimated noise residuals are then input into a weighted averaging method, optimizing PRNU to significantly improve identification performance.

However, further compression during video transmission destroys the PRNU noise components in the video and introduces more complex interference noise, making it difficult to accurately extract PRNU features. Therefore, Video Source Camera Identification (VSCI) presents greater challenges than image-based SCI. To address the various illegal video issues emerging on internet platforms, it is essential to develop VSCI technologies capable of effectively identifying the sources of diverse online videos.

2.3. Identification methods based on deep learning

Methods based on deep learning have shown significant performance improvements in classification tasks recently. In the field of SCI, methods based on deep learning have also demonstrated great potential in VSCI. Li et al. (2024) proposed a multi-level fingerprint learning framework to address the issue of declining accuracy in video integrity and source analysis identification.They integrated video encoding attributes, extracting multilevel features from both decoded video key frames and reference frames. Wang et al. (2024c) utilized integral images to optimize smooth block selection algorithms based on pixel variance, removing interference from semantic video information, and designed a residual neural network with fusion constraint layers to adaptively learn the characteristics of the video source for SCI. Akbari et al. (2024) used their proposed six-stream network to extract low-level and high-level features through the network, and performed camera model identification by fusing features using forward and backward functions based on joint sparse representation. Unlike unimodal methods, Tsingalis et al. (2024) and Dal Cortivo et al. (2021) combined both audio and visual information from videos for VSCI. However, they did not consider that some videos cannot have audio extracted, which limits video identification, therefore necessitating the exploration of more universally applicable methods for VSCI.

It is necessary to develop deep learning methods that primarily learn video visual information containing inherent camera characteristics, making them unaffected by the inability to extract audio from videos, thereby enhancing capabilities for video source camera identification.

3. Proposed method

In this section, we provide an overview of the Dual-branch Hierarchical Feature Fusion Network (DHFFNet) and present a detailed description of its various components.The overview of the model module and its current implementation status are shown in Table 2.

3.1. Overview of the Dual-branch Hierarchical Feature Fusion Network (DHFFNet)

The architecture of DHFFNet is shown in Fig. 2. The model effectively employs dual-branch structure to extract global and local features, and achieves fusion of multi-level features through the Feature Fusion Module (FFM), thereby obtaining accurate video frame classification results. Finally, the model achieves VSCI through majority voting.

3.2. Swin-T network

The Swin Transformer (Swin-T) network (Liu et al., 2021) is an advanced architecture based Vision Transformer. The model employs a hierarchical processing strategy, where the input image is partitioned into multiple non-overlapping patches that undergo progressive transformation across different stages, as illustrated in Fig. 3.

Table 2
Summary of model modules and their implementation status.

Module	Status	Notes
Swin-T	Reused (Liu et al., 2021)	Extract global features
Mutli-stage-CNN	Modified from (Liu et al., 2021)	Extract local features
FFM	Modified from (Huo et al., 2024)	Fusion global and local features
GSAU	Reused (Wang et al., 2024a)	Extract spatial information more effectively
LIA	Reused (Wang et al., 2024b)	Suppress irrelevant noise
CA&SA	Reused (Dai et al., 2017; Hu et al., 2018)	Focus on global or spatial features selectively

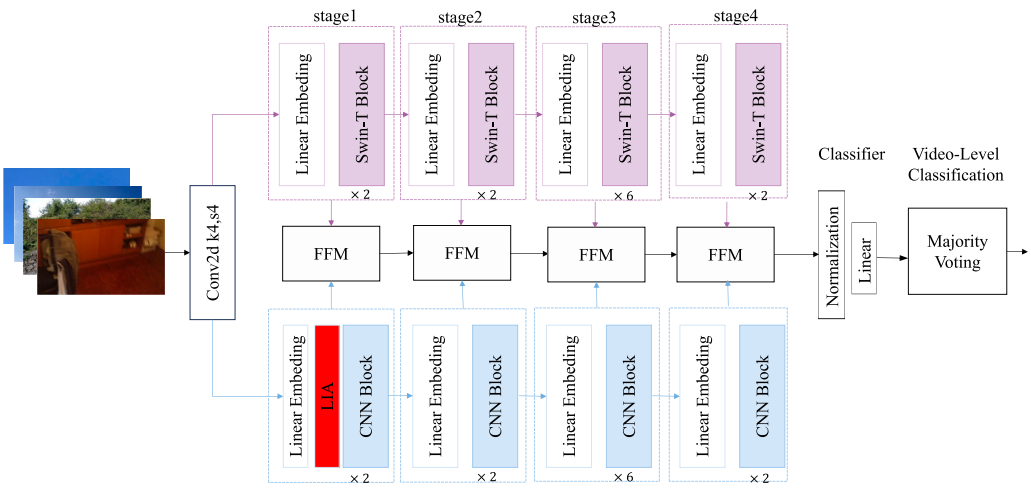

Fig. 2. The overall structure of the Dual-branch Hierarchical Feature Fusion Network (DHFFNet).

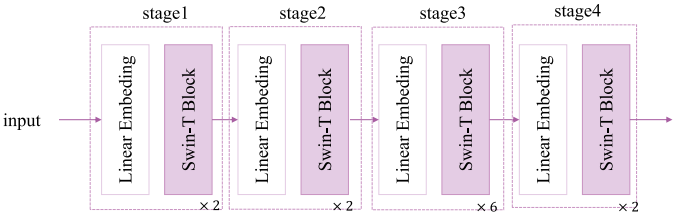

Fig. 3. General architecture of Swin-T network.

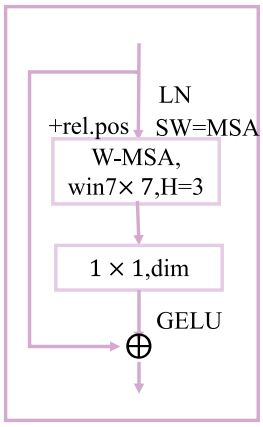

Fig. 4. Swin-T Block architecture.

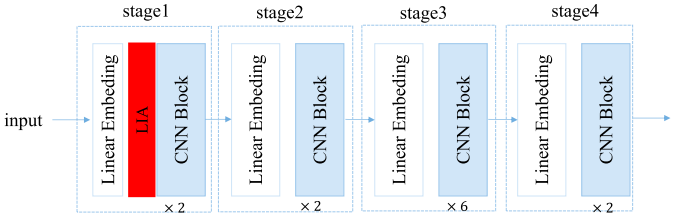

Fig. 5. General architecture of Multi-Stage Convolutional Neural Network.

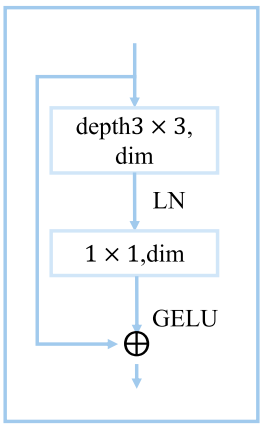

Fig. 6. Convolutional Neural Network block architecture.

As depicted in Fig. 4, the model incorporates a Window Multi-head Self-Attention (W-MSA) mechanism within Swin-T Block at each stage to effectively capture global semantic information (Huo et al., 2024). The computational process of this mechanism is formally expressed in Eq. (1):

$$\begin{aligned} g_i &= f^{1\times1}(W-MSA(LN(G_{i-1}))) + G_{i-1} \\ G_i &= f^{1\times1}(SW-MSA(LN(g_i))) + g_i \end{aligned} \tag{1}$$

where g_i and G_i denote the output features of W-MSA and Shift W-MSA (Wang et al., 2023) operations within the global feature block, respectively. The transformation $f^{1\times1}$ represents a convolution operation with a kernel size of 1×1, and LN signifies the layer normalization operation. The global features are subsequently put into the FFM for further processing.

This architectural design not only significantly reduces computational complexity, but also enhances the accuracy and stability of the model (Lu et al., 2024). However, Transformer has inherent limitations in local feature extraction, which restricts its further application in SCI tasks. Therefore, more effective methods for identification need to be proposed.

3.3. Multi-Stage Convolutional Neural Network

The Multi-Stage CNN specializes in local feature extraction. As shown in Fig. 5, while architecturally similar to Swin-T, it uniquely integrates a Local Importance Attention (LIA) mechanism in the first stage to enhance relevant features and suppress noise. The process employs a 3×3 depth-wise separable convolution (Fig. 6) for efficient local feature extraction, followed by a linear layer enabling cross-channel communication. These processed features are then fed into the feature fusion module, as formulated in Eq. (2):

$$L_i = f^{1\times1}(LN(f^{depth3\times3}(L_{i-1}))) + L_{i-1} \tag{2}$$

where L_i represents the output of Convolutional Neural Network block, $f^{depth3\times3}$ is a depth-wise convolution operation with kernel size 3×3.

3.4. Local Importance-based Attention (LIA)

Prior research typically computes importance maps through subnetworks or matrix operations. Inspired by the local importance acquisition through regional softmax in literature (Gao et al., 2019; Stergiou et al., 2021), Wang et al. (2024b) proposed an attention mechanism based on local importance calculation. Specifically, the importance value of pixel X within the surrounding region R is calculated using Eq. (3):

$$I(X)|_X = \sum_{k\in R}\sum_{i\in R_k} \frac{e^{X_i}}{\sum_{j\in R_k} e^{X_j}} \cdot w_k \tag{3}$$

where $I(X)|_X$ represents the local importance of X. R represents the neighborhood region centered on X. w is a learnable weight used to optimize the calculated importance values. As illustrated in Fig. 7, the implementation employs stacked SoftPool and 3×3 convolutional layers, complemented by stride and squeeze convolutions to enhance computational efficiency and receptive field coverage.

Gating mechanism (Dauphin et al., 2017; Wang et al., 2024a) is used for feature refinement of local importance $L(X)$. To simplify the design, the first channel of the input features is directly selected as the gating signal. Finally, the local importance attention mechanism $L(X)$ can be summarized as Eq. (4):

$$L(X) = \sigma(X_{[0]}) \odot \psi(\sigma(I(X))) \odot X \tag{4}$$

where, $\sigma(.)$ and $\psi(.)$ represent the Sigmoid activation function and bilinear interpolation.

LIA is an attention mechanism, but its applicability in VSCI tasks has not yet been verified. In our task, LIA can adaptively enhance useful features and suppress irrelevant noise through calculating local importance and combining with attention mechanisms, thereby enabling the model to better capture image details and structural information.

3.5. Feature Fusion Module (FFM)

Inspired by Huo et al. (2024), an adaptive feature fusion module (including Channel Attention (CA), Spatial Attention (SA), and Inverted

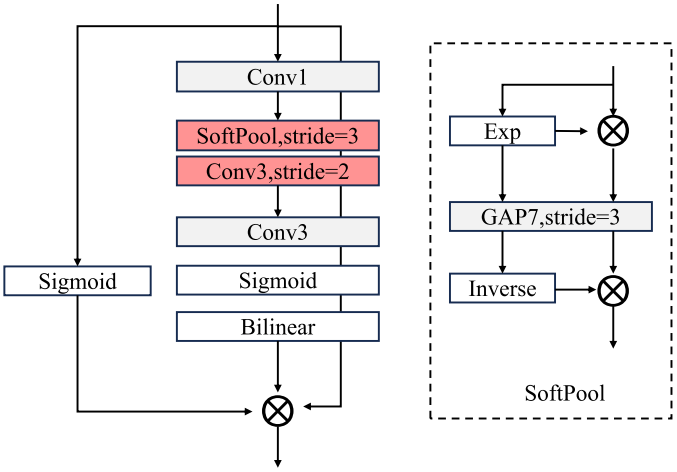

Fig. 7. The architecture of Local Importance-based Attention (LIA).

Residual Multi-layer Perceptron (IRMLP) was employed to fusion global and local features adaptively. Additionally, we introduced Gated Spatial Attention Unit (GSAU) to reduce computational complexity and integrate spatial information. The overall structure of FFM is illustrated in Fig. 8. Here, G_i represents the global features extracted at the current stage, L_i represents the local features extracted at the current stage, F_{i-1} represents the fused features generated by HHB from the previous stage, and F_i represents the fused features generated by HHB at the current stage. The feature fusion operation is shown in Eq. (5):

$$\begin{aligned}
\hat{G}_i &= \mathrm{CA}(G_i) \otimes G_i \\
\hat{L}_i &= \mathrm{SA}(\mathrm{GSAU}(L_i)) \otimes L_i \\
\tilde{F}_i &= \mathrm{Avgpool}(f^{1\times1}(F_{i-1})) \\
\hat{F}_i &= f^{1\times1}(\mathrm{Concat}(G_i, L_i, \tilde{F}_i]) \\
F_i &= \mathrm{IRMLP}(\mathrm{LN}(\mathrm{Concat}[\hat{G}_i, \hat{L}_i, \hat{F}_i])) + \tilde{F}_i
\end{aligned} \tag{5}$$

where $\otimes$ represents element-wise multiplication, $\hat{G}_i$ is generated by CA, and $\hat{L}_i$ is generated by GSAU and SA. $\hat{F}_i$ is the output of the fusion of global features, local features and features from the previous stage. Finally, $\hat{F}_i$, $\hat{G}_i$, and $\hat{L}_i$ are concatenated together, and features are generated through IRMLP.

3.6. Gated Spatial Attention Unit (GSAU)

In Transformer architectures, the Feed-Forward Network (FFN) plays a crucial role as a key component for enhancing feature representation capabilities. However, traditional Multilayer Perceptron (MLP) structures, due to their wide intermediate channel design, often face significant computational burdens when processing large-scale image inputs. To address this challenge, Wang et al. (2024a) inspired by related studies (Chen et al., 2022; Dauphin et al., 2017; Hua et al., 2022; Wang et al., 2022) innovatively combined Spatial Attention (SA) and Gated Linear Unit (GLU) into the proposed GSAU. As shown in Fig. 8, this design not only achieves effective extraction of spatial information but also significantly enhances model performance through the introduction of an adaptive gating mechanism. The core computational process of GSAU can be expressed as Eq. (6):

$$\mathrm{GSAU}(x) = \mathrm{LN}(f_{DW}(x) \otimes scale)) + x \tag{6}$$

where $f_{DW}(.)$ and $\otimes$ indicate depth-wise convolution and element-wise multiplication, respectively, scale represents a learnable parameter for x. GSAU is an adaptive gated attention mechanism, but its applicability in VSCI tasks has not yet been verified. In our task, adding a module with the ability of local feature information interaction can enable the model to achieve more effective extraction of spatial information, thereby improving the performance of our method.

3.7. Channel & spatial attention

The Channel Attention (CA) mechanism enables networks to selectively focus on global features by modeling channel-wise dependencies (Hu et al., 2018), while the Spatial Attention (SA) mechanism emphasizes salient spatial regions (Dai et al., 2017). In the Feature Fusion Module (FFM), global features are processed by the CA mechanism to enhance their discriminative power through channel-wise recalibration. Local features, on the other hand, are refined by the GSAU followed by SA, which jointly amplifies critical spatial regions while suppressing noise, thereby preserving essential local details.

$$\begin{aligned}
\mathrm{CA}(x) &= \sigma(\mathrm{MLP}(\mathrm{AvgPool}(x)) + \mathrm{MLP}(\mathrm{MaxPool}(x))) \\
\mathrm{SA}(x) &= \sigma(f^{7\times7}(\mathrm{Concat}[\mathrm{AvgPool}(x), \mathrm{MaxPool}(x)]))
\end{aligned} \tag{7}$$

where σ represents the Sigmoid function, and $f^{7\times7}$ denotes a convolution operation with a kernel size of 7×7. CA&SA in this paper respectively use enhanced interaction capabilities of global and local feature information, which can achieve better results. Compared with using them alone, it enables the model to learn more comprehensive image information and improves the performance of our method. Compared with using SA and CA for each branch, it can greatly reduce the computing cost and improve the deployment efficiency.

3.8. Video-level predictions

The source camera device classification process for videos employs a robust majority voting mechanism, implemented through the following systematic steps: Initially, 10 frames are randomly sampled from each video to ensure representative coverage. Subsequently, the pre-trained network model is utilized to classify each individual frame, generating prediction results for each sampled frame. Finally, the classification outcomes across all the frames are aggregated, and the device category with the highest frequency is determined as the source camera device of the respective video. This approach significantly improves classification accuracy by leveraging the collective prediction results of multiple frames, thus mitigating the impact of potential outliers or frame-specific anomalies.

Our key contribution lies in the novel integration into a model specifically designed for video source identification. We introduce LIA to redesign the CNN and Swin-T for extracting global and local information, respectively. In the feature fusion module, the incorporation of GSAU and CA&SA enables the model to achieve more effective information extraction, thereby enhancing the performance of our method. This systematic integration has been validated through ablation studies (Section 4.3).

4. Results and analysis

4.1. Datasets

VISION We utilize the Vision dataset (Shullani et al., 2017), which includes images and videos captured under various scenes and imaging conditions. This dataset comprises a total of 35 camera devices, including 29 camera models. According to the dataset splitting strategy, we always retain 80 % of videos from each category (70 % for training set, 10 % for validation set), with the remaining 20 % assigned to the testing set. The labels and detailed information of the dataset used in our experiments are shown in Table 3.

QUFVD The QUFVD dataset (Akbari et al., 2022) consists of 6000 videos captured by 20 devices from 10 brands. Each device is represented by 300 videos, with two devices per model (The QUFVD dataset has already been pre-divided, eliminating the need for manual partitioning.). Table 4 provides detailed information about the QUFVD dataset.

The VISION and QUFVD datasets are very commonly used and well-known in the field of source camera identification. Moreover, corresponding versions uploaded on Facebook and WhatsApp are provided,

Table 3
Details information of VISION dataset.

Brand	Device	Resolution	Label
Apple	iPhone 4S	1080×1920	D02
Apple	iPhone 5c	1080×1920	D05
Apple	iPhone 6	1080×1920	D06
Apple	iPhone 4	1280×720	D09
Apple	iPhone 4S	1080×1920	D10
Apple	iPad 2	1280×720	D13
Apple	iPhone 5c	1080×1920	D14
Apple	iPhone 6	1080×1920	D15
Apple	iPhone 5c	1080×1920	D18
Apple	iPhone 6 Plus	1080×1920	D19
Apple	iPad mini	1080×1920	D20
Apple	iPhone 5	1080×1920	D29
Apple	iPhone 5	1080×1920	D34
Samsung	Galaxy S III Mini GT-I8190N	1280×720	D01
Samsung	Galaxy GT-P5210	1280×720	D08
Samsung	Galaxy S3 GT-I9300	1080×1920	D11
Samsung	GalaxyTrendPlusGT-S7580	1280×720	D22
Samsung	Galaxy S III Mini GT-I8190	1280×720	D26
Samsung	Galaxy S5 SM-G900F	1080×1920	D27
Samsung	Galaxy S4 Mini GT-I9195	1080×1920	D31
Samsung	Galaxy Tab A SM-T555	1280×720	D35
Huawei	P9 EVA-L09	1080×1920	D03
Huawei	P9 Lite VNS-L31	1080×1920	D16
Huawei	P8 GRA-L09	1080×1920	D28
Huawei	Honor 5C NEM-L51	1080×1920	D30
Huawei	Ascend G6-U10	1280×720	D33
LG electronics	D290	800×480	D04
Lenovo	P70-A	1280×720	D07
Sony	Xperia Z1 Compact-D5503	1280×720	D12
Microsoft	Lumia 640LTE	1280×720	D17
Wiko	Ridge 4G	1280×720	D21
Asus	Zenfone 2 Laser	640×480	D23
Xiaomi	Redmi Note 3	1280×720	D24
OnePlus	A3000	1280×720	D25
OnePlus	A3003	1280×720	D32

Table 4
Details information of QUFVD dataset.

Brand	Device	Resolution	Label
Samsung	Galaxy A50-1	1080×1920	D1
Samsung	Galaxy A50-2	1080×1920	D2
Samsung	Note9-1	1080×1920	D3
Samsung	Note9-2	1080×1920	D4
Huawei	Y7-1	720×1280	D5
Huawei	Y7-2	720×1280	D6
Huawei	Y9-1	720×1280	D7
Huawei	Y9-2	720×1280	D8
iPhone	8 Plus-1	1080×1920	D9
iPhone	8 Plus-2	1080×1920	D10
iPhone	XS Max-1	1080×1920	D11
iPhone	XS Max-2	1080×1920	D12
Nokia	5.4-1	1080×1920	D13
Nokia	5.4-2	1080×1920	D14
Nokia	7.1-1	1080×1920	D15
Nokia	7.1-2	1080×1920	D16
Xiaomi	Redmi Note8-1	1080×1920	D17
Xiaomi	Redmi Note8-2	1080×1920	D18
Xiaomi	Redmi Note9 Pro-1	1080×1920	D19
Xiaomi	Redmi Note9 Pro-2	1080×1920	D20

and the collected content is also exchanged through social media platforms, taking into account the variability in the real world (such as editing, compression, and resolution differences).

4.2. Experiment details and evaluation metrics

We implement our PyTorch-based method by training on an NVIDIA RTX 2080 GPU. The training is conducted for a total of 100 epochs, with a base learning rate of 1e-4 and batch size of 16. For consistency, this paper adopts a uniform video frame size of 224×224. To maintain consistency, we use the same training, validation and testing sets as in previous works. We select Accuracy (ACC) as the classification metric. The computational complexity of the model is relatively high. As shown in Table 5, the FLOPs is 18.66G, the Parameters are 125.74M, and the classification time of one frame during classification is 5.62 ms. The model is conducted under the framework of Contributors (2018), utilizing the classification cross-entropy loss function to calculate the loss:

$$CrossEntropyLoss = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{k} y_i^l \log y_i^p \tag{8}$$

Table 5
DHFFNet specific parameters.

Parameter	Number
FLOPs	18.66G
Parameters	125.74M
Time for classifying each frame	5.62 ms

Table 6
Compared with SOTA methods on VISION dataset at model level.

Method	ACC (%)
LHFMF (Li et al., 2024)	97.19
GCD (Korgialas et al., 2024)	97.07
CMMCMI (Dal Cortivo et al., 2021)	99.00
CMIAVC (Tsingalis et al., 2024)	95.38
Ours	99.20

Table 7
Experiments at the device and model level across VISION and QUFVD datasets.

		ACC(%)	
Dataset	Level	Frame level	Video level
VISION	29 Cameras (model level)	97.88	99.20
	35 Cameras (device level)	97.68	98.94
QUFVD	10 Cameras (model level)	84.41	85.00
	20 Cameras (device level)	64.33	65.25

4.3. Experimental results and analysis

Comparison with the state-of-the-art systems: We conducted comparative experiments with current SOTA methods to validate the superiority of our method, and the results are shown in Table 6. Our proposed method, which integrates global and local features, achieved an identification accuracy of 99.20 %, outperforming existing SOTA methods in feature fusion. The experimental results prove the effectiveness of the dual-branch hierarchical feature fusion strategy. By integrating global and local features, our model can more comprehensively capture the unique features of each camera model in video data.

Performance on other datasets: As shown in Table 7, we conducted comparative experiments on both the VISION and QUFVD datasets to comprehensively evaluate our method's performance. On the QUFVD dataset, our method achieved identification accuracies of 85.33 % at the model level and 64.67 % at the device level. In contrast, on the VISION dataset, it attained 99.20 % and 98.94 % at model and device levels, respectively.

These results demonstrate that our model achieves superior accuracy on the VISION dataset. At the model level, our proposed approach also delivers satisfactory performance on the QUFVD dataset, though its overall accuracy remains slightly lower than that achieved on the visual dataset. This discrepancy primarily stems from the substantial volume fluctuations in QUFVD's data distribution, which lead to severe class imbalance. The limited training samples for certain categories further reduce identification accuracy.

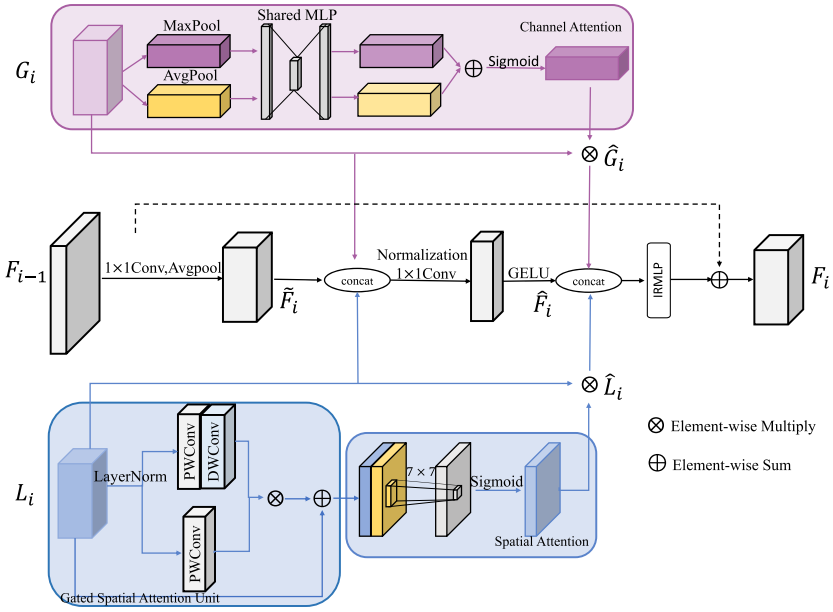

Fig. 8. The architecture of feature fusion module.

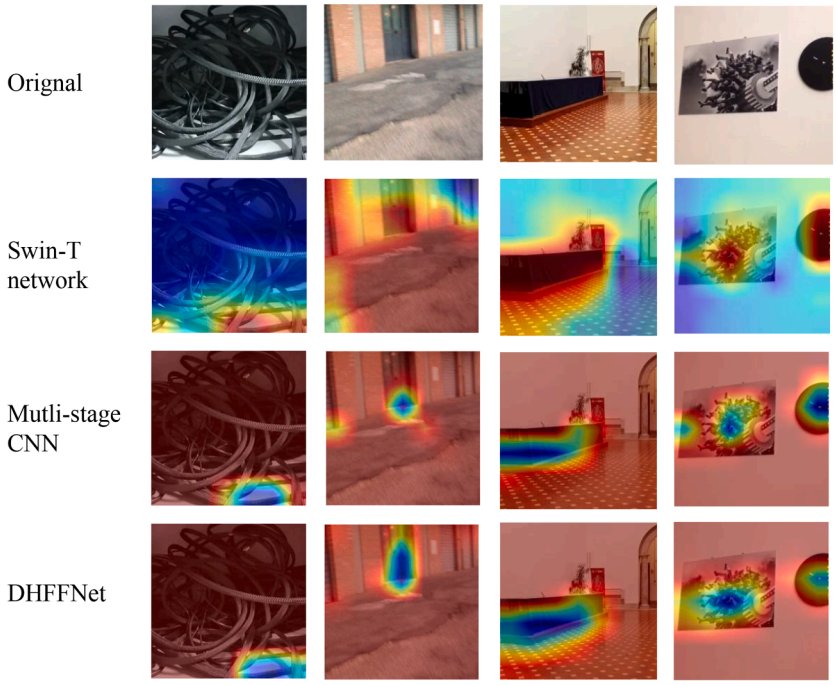

Fig. 9. Grad-CAM visualizations.

Additionally, while both VISION and QUFVD are widely adopted benchmarks, existing models (including ours) have not been rigorously evaluated on strictly edited or AI-generated videos. Future work will focus on improving robustness against synthetic content, cross-dataset generalization, and mitigating data scarcity to enhance source identification performance.

The confusion matrices for the model on the VISION and QUFVD datasets are present in Fig. 10–13. In particular, the accuracy of the model source identification is higher than the accuracy of the device source identification. Among these cases, the majority of misclassified device models belong to the same brand. For example, in the QUFVD model source identification experiment, our method incorrectly classified 13 % of the original videos captured by the iPhone XS Max as the iPhone 8 Plus. This shows that video processing workflows may erode subtle distinctive traces between different device models, leading to increased misidentification rates among devices from the same brand.

Ablation study: We evaluated the impact of various components on model performance using the QUFVD dataset, and the experimental results are shown in Table 8. The experimental results indicate that the introduction of the local branch significantly improved model performance, with frame-level and video-level identification accuracy increasing by 7.76 % and 5.92 % respectively. As shown in Fig. 9, this performance improvement mainly benefits from the synergistic effect of the dual-branch network structure: the global branch captures overall features while the local branch extracts detailed information, complementing each other to enhance the model's representation capability.

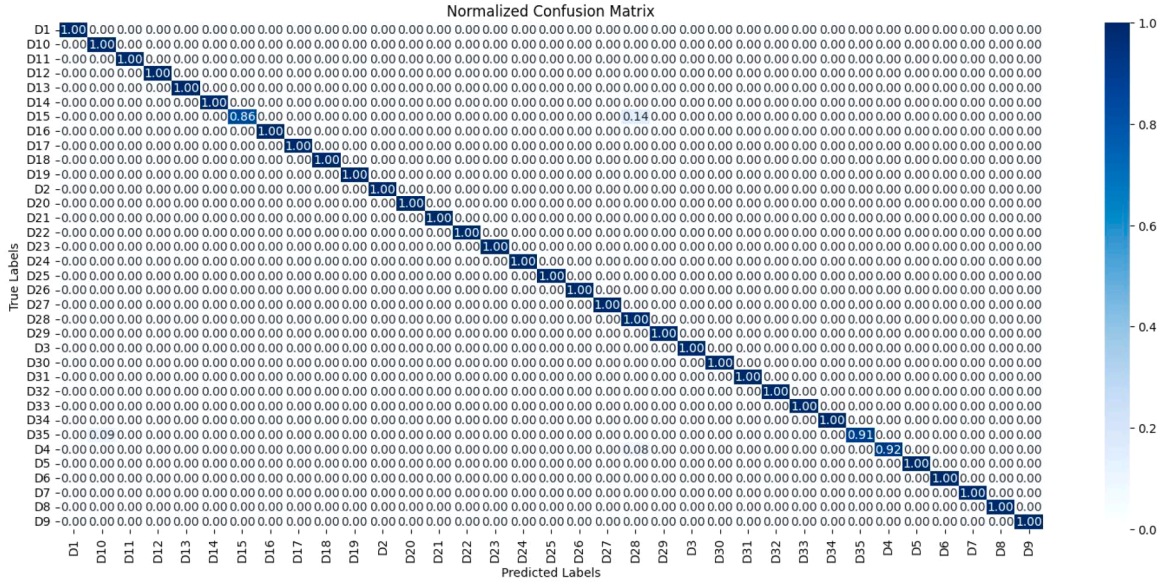

Fig. 10. Comparative experimental confusion matrixs on VISION dataset at device level.

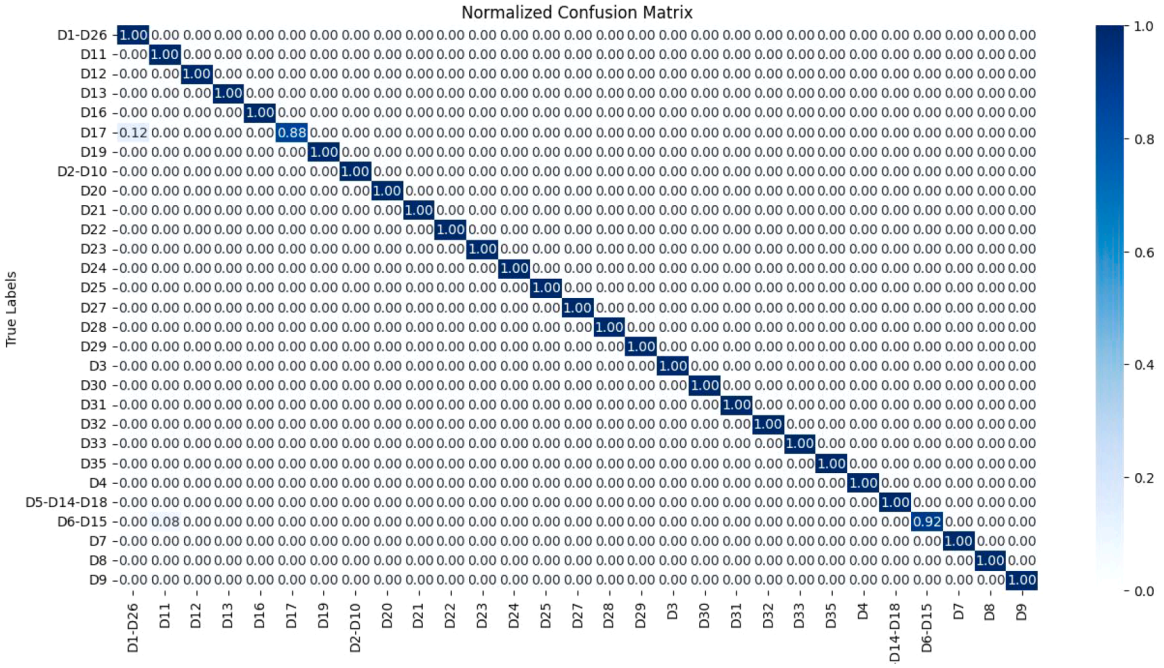

Fig. 11. Comparative experimental confusion matrixs on VISION dataset at model level.

Table 8
Ablation experiments on QUFVD dataset at device level.

Method	ACC (%)	
	Frame level	Video level
Swin-T network	55.54	57.83
+ Mutli-stage CNN	63.30	63.75
+ LIA	63.39	64.08
+ SA&CA	63.77	64.83
+ GSAU	64.33	65.25

The LIA mechanism, through combining local importance calculation with attention mechanisms, adaptively enhances useful features and suppresses irrelevant noise, enabling the model to more precisely capture image details and structural information. Although LIA is a local attention mechanism, it successfully achieves higher-order information interaction effects comparable to global attention through local importance modeling and gating mechanisms. CA selectively enhances important channel features, while SA focuses on key spatial regions. After introducing CA and SA mechanisms in the FFM, the classification accuracy is further improved, achieving effective fusion of global and local features.

GSAU effectively integrates spatial information by introducing gating mechanisms and Spatial Attention. After adding GSAU, the performance of the model improved by 0.42 %, proving that GSAU can improve the performance of the model.

Through ablation experiments by gradually removing or adding modules, the experiments show that each module (Local Feature Extract Path, CA&SA, LIA, and GSAU contributes to performance improvement.

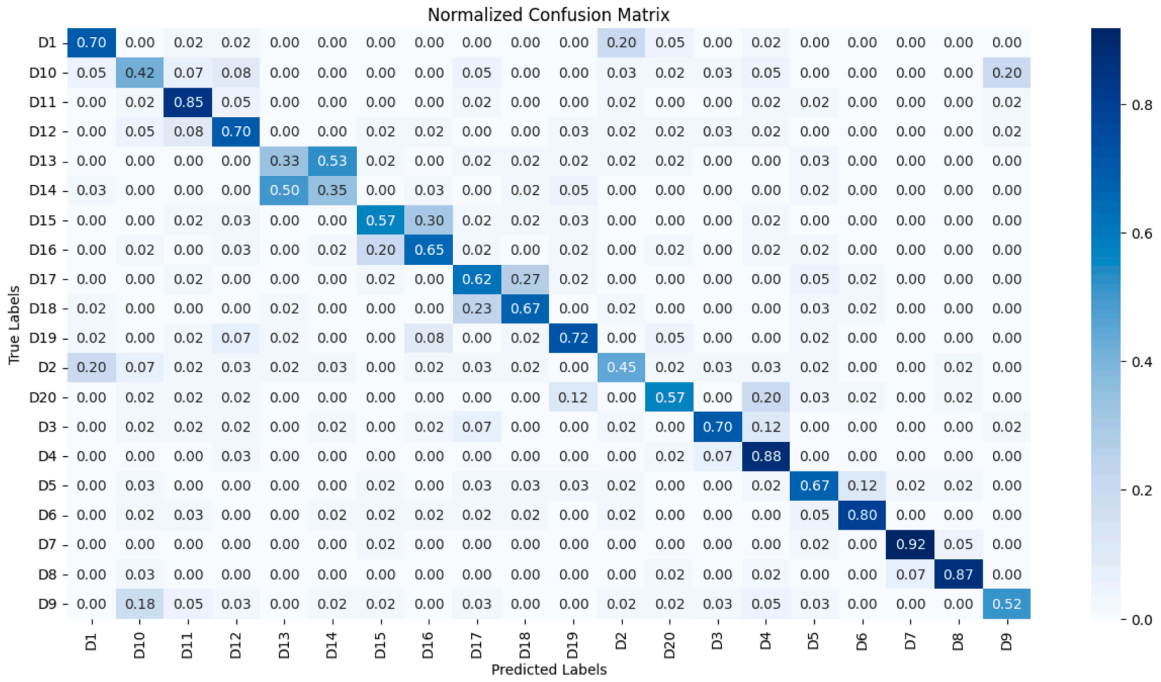

Fig. 12. Comparative experimental confusion matrixs on QUFVD dataset at device level.

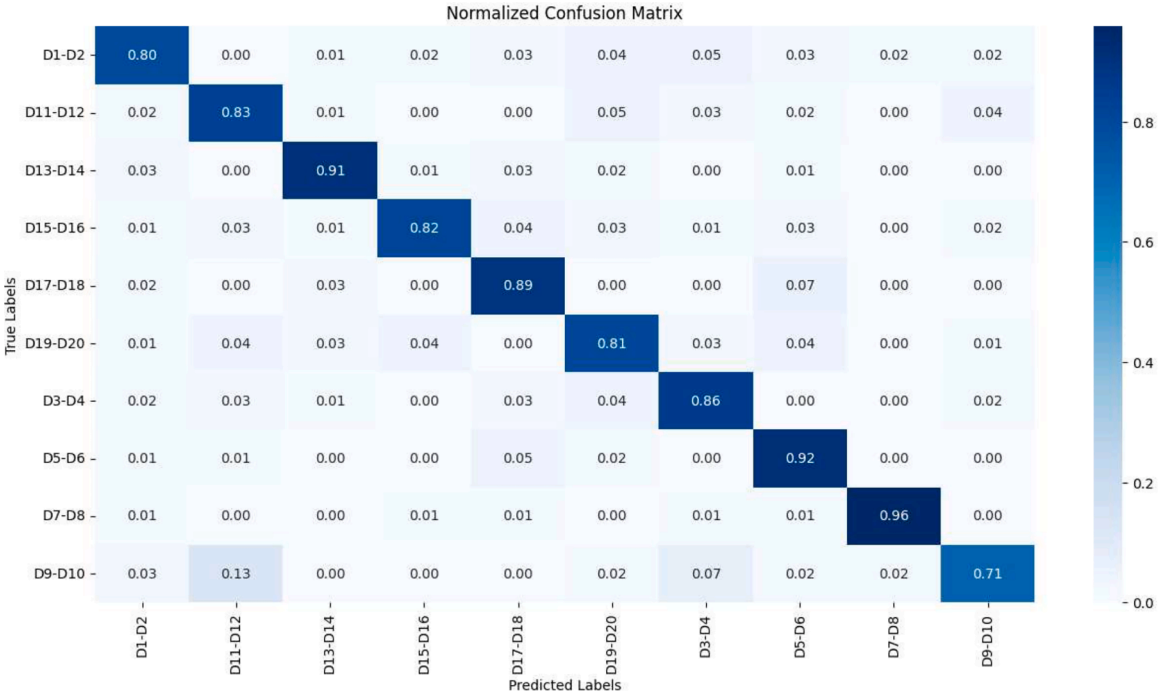

Fig. 13. Comparative experimental confusion matrixs on QUFVD dataset at model level.

Table 9
Sampling sensitivity analysis on VISION and QUFVD dataset at device level.

Acc and dataset	VISION					QUFVD				
	1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th
ACC (%)	98.94	98.67	99.20	98.67	99.20	65.17	65.33	65.25	65.50	65.00
Average ACC (%)			98.94					65.25		

Sampling sensitivity analysis: For the VISION dataset, the accuracy across 5 trials ranged between 98.67 % and 99.20 %, with a standard deviation of 0.23 %. The minimal variation confirms that random sampling introduces negligible bias for this dataset, likely due to its high inter-frame consistency. The QUFVD dataset showed slightly greater variation in experimental accuracy (65.00 %–65.50 %, std: 0.19 %), indicating marginally higher sensitivity to frame selection. This may stem from QUFVD's more diverse intra-video content, though the limited range ($\leq 0.5\,\%$) still demonstrates robustness. Our experimental results demonstrate that 10-frame sampling achieves stable performance for both datasets. The specific experimental data can be seen in Table 9.

5. Conclusion

In this paper, we propose a novel Dual-branch Hierarchical Feature Fusion Network (DHFFNet) for video camera source identification. The model extracts global and local features through a dual-branch architecture, where the Multi-Stage Convolutional Neural Network (MSCNN) incorporates a Local Importance-based Attention (LIA) mechanism to adaptively enhance local information. These features are then fused by a dedicated feature fusion module empowered by Gated Spatial Attention Unit (GSAU), Spatial Attention (SA), and Channel Attention (CA) mechanisms. During inference, the model classifies individual frames, aggregates the predictions via majority voting, and assigns the most frequent device category as the final video-level source. The experimental results demonstrate that DHFFNet outperforms the comparative methods on both VISION and QUFVD datasets.

However, limitations remain in handling small-sample datasets and videos from similar sensor devices, and the robustness against heavily edited or AI-generated videos requires further evaluation. Future work will focus on: (1) advancing device-level VSCI, (2) optimizing model deployment for mobile/edge platforms, and (3) evaluating adversarial robustness.

CRediT authorship contribution statement

Bo Wang: Conceptualization, Methodology, Software, Formal analysis, Validation, Visualization, Writing – original draft, Writing – review & editing; **Jiaqi Chi:** Methodology, Formal analysis, Visualization, Writing – review & editing; **Zhuocheng Wu:** Conceptualization, Methodology, Formal analysis, Writing – review & editing; **Huimin Liu:** Conceptualization, Methodology, Writing – review & editing; **Wei Wang:** Conceptualization, Methodology, Writing – review & editing, Supervision, Resources.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Akbari, Y., Al-Maadeed, S., Al-Maadeed, N., Al-Ali, A., Khelifi, F., Lawgaly, A. et al. (2022). A new forensic video database for source smartphone identification: Description and analysis. *IEEE Access*, *10*, 20080–20091.

Akbari, Y., Al Maadeed, S., Elharrouss, O., Ottakath, N., & Khelifi, F. (2024). Hierarchical deep learning approach using fusion layer for source camera model identification based on video taken by smartphone. *Expert Systems with Applications*, *238*, 121603.

Anmol, T., & Sitara, K. (2024). Video source camera identification using fusion of texture features and noise fingerprint. *Forensic Science International: Digital Investigation*, *49*, 301746.

Bennabhaktula, G. S., Timmerman, D., Alegre, E., & Azzopardi, G. (2022). Source camera device identification from videos. *SN Computer Science*, *3*(4), 316.

Berdich, A., Groza, B., & Mayrhofer, R. (2023). A survey on fingerprinting technologies for smartphones based on embedded transducers. *IEEE Internet of Things Journal*, *10*(16), 14646–14670.

Chen, L., Chu, X., Zhang, X., & Sun, J. (2022). Simple baselines for image restoration. In *European conference on computer vision* (pp. 17–33). Springer.

Contributors, M. (2018). Mmcv: Openmmlab computer vision foundation.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 764–773).

Dal Cortivo, D., Mandelli, S., Bestagini, P., & Tubaro, S. (2021). Cnn-based multi-modal camera model identification on video sequences. *Journal of Imaging*, *7*(8), 135.

Dauphin, Y. N., Fan, A., Auli, M., & Grangier, D. (2017). Language modeling with gated convolutional networks. In *International conference on machine learning* (pp. 933–941). PMLR.

Elharrouss, O., Akbari, Y., Almadeed, N., Al-Maadeed, S., Khelifi, F., & Bouridane, A. (2025). Pdc-vit: Source camera identification using pixel difference convolution and vision transformer. *Neural Computing and Applications*, *37*, 6933–6949.

Gao, Z., Wang, L., & Wu, G. (2019). Lip: Local importance-based pooling. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3355–3364).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Hua, W., Dai, Z., Liu, H., & Le, Q. (2022). Transformer quality in linear time. In *International conference on machine learning* (pp. 9099–9117). PMLR.

Huo, X., Sun, G., Tian, S., Wang, Y., Yu, L., Long, J., Zhang, W., & Li, A. (2024). Hifuse: Hierarchical multi-scale feature fusion network for medical image classification. *Biomedical Signal Processing and Control*, *87*, 105534.

Korgialas, C., Tzolopoulos, G., & Kotropoulos, C. (2024). On explainable closed-set source device identification using log-mel spectrograms from videos' audio: A grad-CAM approach. *IEEE Access*, *12*, 121822–121836.

Kurosawa, K., Kuroki, K., & Saitoh, N. (1999). Ccd fingerprint method-identification of a video camera from videotaped images. In *Proceedings 1999 international conference on image processing (cat. 99CH36348)* (pp. 537–540). IEEE (*vol. 3*).

Lawgaly, A., Khelifi, F., Bouridane, A., Al-Maaddeed, S., & Akbari, Y. (2022). Prnu estimation based on weighted averaging for source smartphone video identification. In *2022 8th International conference on control, decision and information technologies (coDIT)* (pp. 75–80). IEEE (*vol. 1*).

Li, Y., Ye, J., Zeng, L., Liang, R., Zheng, X., Sun, W., & Wang, N. (2024). Learning hierarchical fingerprints via multi-level fusion for video integrity and source analysis. *IEEE Transactions on Consumer Electronics*, *70*(1), 3414–3424.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022).

Lu, J., Li, C., Huang, X., Cui, C., & Emam, M. (2024). Source camera identification algorithm based on multi-scale feature fusion. *Computers, Materials & Continua*, *80*(2), 3047–3065.

Manisha, Li, C. T., & Kotegar, K. A. (2023). Source camera identification with a robust device fingerprint: evolution from image-based to video-based approaches. *Sensors*, *23*(17), 7385.

Ngharamike, E., Ang, L. M., Phooi, S. K., & Wang, M. (2023). Exploiting the rolling shutter read-out time for ENF-based camera identification. *Applied Sciences*, *13*(8), 5039.

Rana, K., Singh, G., & Goyal, P. (2023). Snrcn2: Steganalysis noise residuals based cnn for source social network identification of digital images. *Pattern Recognition Letters*, *171*, 124–130.

Shullani, D., Fontani, M., Iuliani, M., Shaya, O. A., & Piva, A. (2017). Vision: A video and image dataset for source identification. *EURASIP Journal on Information Security*, *2017*(1), 15.

Stergiou, A., Poppe, R., & Kalliatakis, G. (2021). Refining activation downsampling with softpool. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10357–10366).

Sychandran, C. S., & Shreelekshmi, R. (2024). Sccrnet: A framework for source camera identification on digital images. *Neural Computing and Applications*, *36*(3), 1167–1179.

Tsingalis, I., Korgialas, C., & Kotropoulos, C. (2024). Camera model identification using audio and visual content from videos. *arXiv preprint arXiv:2406.17916*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 1–11.

Wang, S., Gao, Z., & Liu, D. (2023). Swin-GAN: Generative adversarial network based on shifted windows transformer architecture for image generation. *The Visual Computer*, *39*(12), 6085–6095.

Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2022). Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, *8*(3), 415–424.

Wang, Y., Li, Y., Wang, G., & Liu, X. (2024a). Multi-scale attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5950–5960).

Wang, Y., Li, Y., Wang, G., & Liu, X. (2024b). PlainUSR: Chasing faster convnet for efficient super-resolution. In *Proceedings of the Asian conference on computer vision* (pp. 4262–4279).

Wang, Y., Sun, Q., & Rong, D. (2024c). Generalizing source camera identification based on integral image optimization and constrained neural network. *Electronics*, *13*(18), 3630.

Yang, P., Baracchi, D., Iuliani, M., Shullani, D., Ni, R., Zhao, Y., & Piva, A. (2020). Efficient video integrity analysis through container characterization. *IEEE Journal of Selected Topics in Signal Processing*, *14*(5), 947–954.

Yang, W. C., Jiang, J., & Chen, C. H. (2021). A fast source camera identification and verification method based on PRNU analysis for use in video forensic investigations. *Multimedia Tools and Applications*, *80*(5), 6617–6638.