Knowledge-Based Systems 351 (2026) 116444

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Siameformer: Towards a robust source camera identification method on lossy images

Bo Wang [a], Jiaqi Chi [a], Weiming Zheng [a], Fei Wei [b], Yi Li [c,*]

[a] School of Information and Communication Engineering, Dalian University of Technology, No. 2 Linggong Road, Ganjingzi District, Dalian, 116024, Liaoning, China
[b] Alibaba Group, No. 969 WenYiXi Street, Yuhang District, Hangzhou, 311121, Zhejiang, China
[c] School of Artificial Intelligence, Dalian University of Technology, No. 2 Linggong Road, Ganjingzi District, Dalian, 116024, Liaoning, China

ARTICLE INFO

Keywords:
Source camera identification
Multi-head attention
Siamese network
Vision transformer

ABSTRACT

Source camera identification (SCI) is a common problem in digital forensics, aiming to identify the source device of given images. Traditional digital image source forensics are primarily conducted in original scenarios. However, in real-world scenarios, digital images disseminated through social networks are often subjected to various complex interferences, leading to a significant decline in the performance of existing image source forensic methods. To address the issue of network image source forensics in real-world scenarios, we propose a network image source forensics method called **Siameformer**. This method integrates a multi-head attention-based Siamese network with convolutional neural networks (CNN) and Vision Transformers (ViT) for image source identification. We designed comprehensive experiments, and our model achieved satisfactory performance on benchmark public databases (such as Dresden, VISION, and Forchheim).

1. Introduction

Due to the prevalence of various social media platforms,people can easily and quickly share images captured by themselves or others through the internet in their daily lives. However, this convenience comes with serious security issues, such as image infringement, video tampering, and the posting of illegal or criminal content. The anonymity mechanism of the internet further increases the possibility for these criminals to evade punishment. Therefore, there is an urgent need for a fast, reliable, and robust technology to verify the source and authenticity of multimedia information. Source camera identification (SCI), a technique focused on identifying the characteristics or identity of the image source, helps ensure authenticity, combat misinformation, support forensic investigations, protect copyrights, and enhance digital privacy and security.

Source camera identification (SCI) is generally categorized into two levels of device attribution: (1) camera model identification and (2) specific device identification. In practical forensic scenarios, messaging applications, or other compressed transmission channels. The available images are often degraded by operations such as lossy compression, resizing, and requantization. While it is possible to collect original, unaltered images under controlled laboratory conditions to train forensic classifiers, such pristine data is seldom accessible in real-world investigations, where only transmitted versions of images are typically available. A key challenge in real-world source camera identification, therefore, lies in performing reliable attribution when only a degraded, network-transmitted version of an image is accessible, rather than its original, high-quality counterpart. Existing forensic methods [1–3], when applied directly to such interfered network images, often suffer from a noticeable decline in accuracy, as critical device-specific traces are distorted or removed during transmission. As illustrated in Fig. 1, we contrast two forensic scenarios: one using original images and the other using network-degraded images, highlighting the performance gap introduced by image degradation in operational settings.

In traditional machine learning algorithms, [4] analyzed the impact of JPEG lossy compression on algorithm robustness from a feature perspective. They proposed an optimization method to enhance robustness by extracting image fingerprint features based on noise differences between distorted and original images. Building on this foundation, [5] restored distorted images using a mosaic removal algorithm. By comparing residual information between original and restored images, they extracted mosaic removal residual features for robust source identification. In the realm of deep learning-based methods for network image source identification, [6,7] directly learned from distorted images to

* Corresponding author.
E-mail addresses: bowang@dlut.edu.cn (B. Wang), 0703chi@mail.dlut.edu.cn (J. Chi), dutzhengweiming@163.com (W. Zheng), feiwei@alibaba-inc.com (F. Wei), liyi@dlut.edu.cn (Y. Li).

https://doi.org/10.1016/j.knosys.2026.116444
Received 16 January 2025; Received in revised form 1 February 2026; Accepted 12 June 2026
Available online 3 July 2026

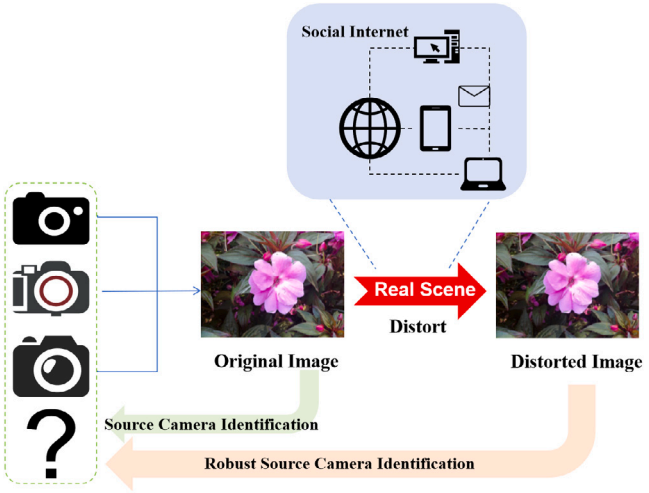


Fig. 1. Source camera identification in both original and real scene settings.

improve method robustness. Zhang et al. [8] introduced a tripartite transfer learning model to enhance robustness through transfer learning.

Based on swing-t networks, [9] use Transformer Blocks and Graph Convolutional network (GCN) modules to fuse multi-scale features from different stages of the backbone network. Jaiswal and Srivastava [10] proposes a framework based on frequency and spatial features that uses discrete wavelet transform (DWT) and local binary mode (LBP) to extract these features from images and train them on multi-class classifiers such as support vector machines (SVM), linear discriminant analysis (LDA), and K-nearest neighbor (KNN).Sychandran and Shreelekshmi [11]proposes a new architecture that combines convolutional layers and residual blocks to distinguish source cameras, helping to learn and extract features used to recognize models and sensor-level patterns. Although convolutional neural networks (CNNS) perform well in local feature extraction, their ability to capture global information is limited, which limits their recognition performance in complex scenes. Transformer has excellent ability to model global semantic information [12]. However, Transformers has inherent limitations in local feature extraction, which restricts its further application in SCI tasks.

Dong et al. [13] proposed that CSWin Transformer contains a cross-shaped window self-attention mechanism and convolution based locally enhanced position coding, which shows excellent performance in visual tasks.Li et al. [14] introduced the features of convolution into ViT to realize efficient representation learning by simulating separable convolution operations with self-attention mechanism. However, it can be seen from the above studies that most of the studies combining the two are in the mixed mode of “one master and one pair”. That is, one mechanism is dominant and another mechanism is auxiliary to it. However, both the convolution operation of CNN and the self-attention mechanism of ViT have their own advantages, and the mixed mode of “one master and one pair” obviously cannot fully release and combine the potential of the two.

To tackle these challenges, we propose a new approach using a siamese network model. The self-attention mechanism is primarily used to capture the dependencies between different positions in the input feature sequence, allowing it to better capture the global information of feature vectors and overcome the locality limitation of Convolutional Neural Networks (CNNs). As the core mechanism of the Transformer neural network architecture, self-attention has gradually been extended to fields such as computer vision, leading to the development of corresponding algorithmic models like the Vision Transformer (ViT). Therefore, combining self-attention-based ViTs with CNNs to complement their respective shortcomings is a feasible solution.

Our model integrates Convolutional Neural Networks (CNN) and Vision Transformer (ViT) with a multi-head attention mechanism, harnessing their combined power. We achieve this by adaptively extracting image features using a combination of multi-head attention and convolution. The self-attention mechanism is primarily used to capture the dependencies between different positions in the input feature sequence, enabling it to effectively handle various types of input sequences. This allows us to effectively capture the relevant information for source identification. Furthermore, we integrate both surface features and deep features using a local/global attention module. This helps us to leverage the strengths of both types of features, leading to more accurate identification results. To optimize the performance of our model and enhance its robustness, we introduce a siamese network weight sharing mechanism. This mechanism emphasizes the fingerprint features of the image, improving the model’s ability to handle different sources and variations. By incorporating these innovations, our approach offers an advanced solution for image source identification, enhancing both accuracy and robustness, which is justified by our comprehensive experiments.

In this paper, we propose the **Siameformer** method to solve the problem of forensic models lacking robustness in real-world scenarios. Following is a summary of the main contributions of this paper:

- We propose an extended two-branch structure that incorporates different types of feature domains to capture a wider range of information, thereby improving identification accuracy.
- By leveraging a self-attention mechanism, we integrate convolutional neural networks (CNN) and Vision Transformers (ViT) to facilitate feature extraction guided by global information.
- To improve the model’s performance, we employ a Siamese network architecture with shared weight parameters, enabling the training of a residual graph for classification (depicted in Fig. 2).

The paper is structured as follows. Section 2 presents a review of available methods for source camera classification. Section 3 introduces the proposed network architecture. In Section 4, we present the validity and robustness of the proposed method for source forensics of network images in real scene on popular datasets. Finally, we conclude the paper with Section 5.

2. Related work

In the traditional original scene setting, SCI methods mainly extract discriminative features for source identification. The most classic SCI method was proposed by Lukas et al. [15]. During imaging, camera sensors attach artifacts to the output image, known as Sensor Pattern Noise (SPN) [15]. SPN has two main components: Dark Signal Non-Uniformity (DSNU) and Photo-Response Non-Uniformity (PRNU). DSNU refers to the uneven response of pixels to the environment when the sensor is not exposed to light. Therefore, it is also called dark current noise. Under normal lighting conditions, this signal is relatively weak in the image. Since most of the sensor pattern noise consists of PRNU noise components, sensor pattern noise has also come to be known as PRNU noise. Photo Response Non-Uniformity (PRNU), caused by inherent sensor hardware defects in imaging devices, is an exceptionally unique component of pattern noise. It was one of the earliest discovered and remains the most commonly used fingerprint feature for individual device source identification in image forensics. Later, PRNU was further studied as a key feature of SCI [16–20].

More features were subsequently discovered and applied to source identification [21]. Based on the color and quality differences of images taken by different cameras, [22] constructed a 34-dimensional feature set including image color, quality, and higher-order baud sign for source identification. Bayram et al. [23] studied the CFA interpolation process during camera shooting and traced the source through CFA interpolation coefficients. In addition to PRNU and CFA interpolation coefficients, hardware features (hardware correlation feature [24], camera white balance [25], lens radial distortion [26], periodic joint noise

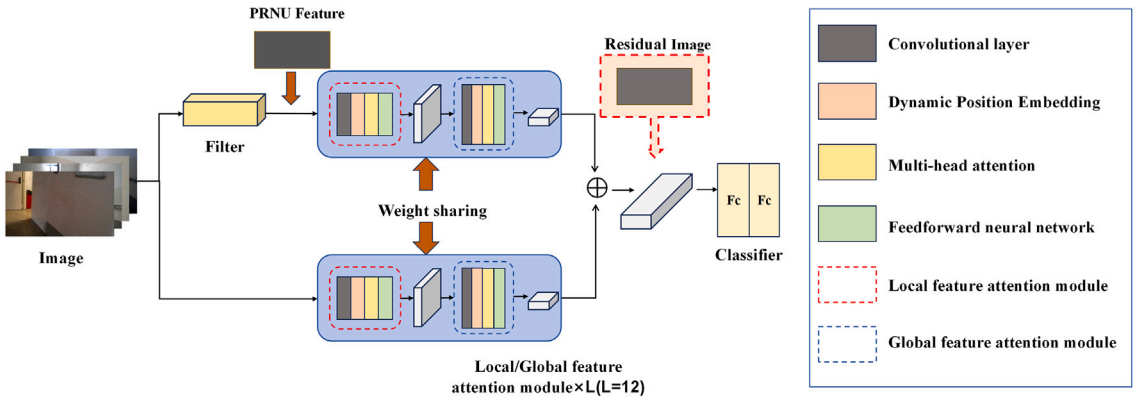


Fig. 2. The methodological framework of **Siameformer**. The upper and lower branches are connected by weight sharing mechanism to enable residual training.

NUA [27]), software features (image Local Binary Pattern (LBP) [28], color correlation [29]) and their joint features [1] have also achieved high accuracy in this classification task. Goljan et al. [4] analyzed the influence of JPEG lossy compression on the robustness of the algorithm from the feature perspective and proposed an optimization method to extract image fingerprint features according to the noise difference between the distorted image and the original image to improve robustness. Based on this, [5] used a Mosaic removal algorithm to restore the distorted image. By comparing the residual information between the original image and the recovered image, a residual removal feature of the image Mosaic is extracted for robust source identification.

Traditional SCI methods based on hand-crafted feature extraction are limited and face difficulties in extracting effective features. The development of deep learning techniques has provided a new approach, using neural networks to automatically extract features from images. Baroffio et al. [2] introduced deep learning algorithms into the SCI field for the first time and addressed source forensics problems using AlexNet. However, simple transfer models perform poorly when dealing with multi-class datasets. Tuama et al. [30] optimized the network model for SCI problems by adding a pooling layer and residual checks. Yang et al. [31] proposed a content-adaptive fusion residual network, which selects different ResNet branches based on image content. Chen et al. [32] combined ResNet with support vector machines to achieve effective classification. Densenet [33], CCN-net [34] and VGG-net [35], known for their superior performance, have also been introduced into SCI. David et al. [36] proposed a simple CNN model based on mobile phone image characteristics. The robustness of SCI algorithms based on deep learning also faces challenges. Lin et al. [7] studied interference images directly to improve method robustness. By introducing transfer learning, [8] proposed a ternary transfer learning model to enhance method robustness. Wang et al. [37] proposed an SCI method based on adversarial training to counteract malicious fine image modifications, thereby improving model robustness against attacks.

Promoting only image features or network models is an effective strategy, but it is not enough. The focus of this paper is on integrating the features of interfered images with the advantages of network models to propose a more effective method for robust source identification. It is worth exploring how to adaptively extract undisturbed local features for source identification based on the correlation between image content and features, leveraging CNN and ViT to effectively capture both local and global features, ensuring robustness.

3. Siameformer mechanism

For conducting source identification on distorted images, we introduce the **Siameformer**, a robust SCI mechanism combining Convolutional Neural Networks (CNN) and Vision Transformer (ViT), as shown in Fig. 2. The structure of the network proposed is presented in the figure. The succeeding subsection details structural components of the approach implemented in this paper.

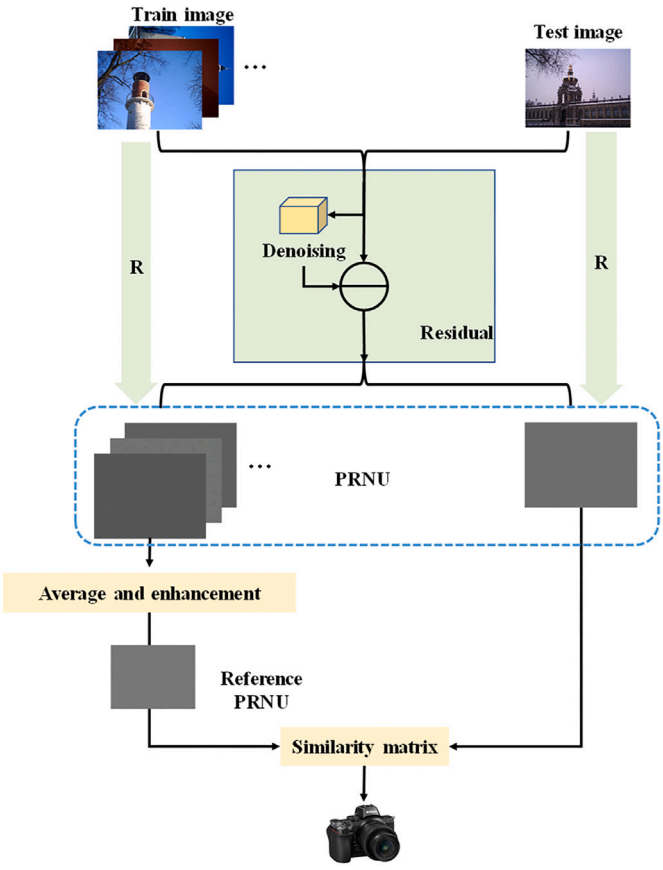


Fig. 3. The basic framework of SCI is based on PRNU noise feature. The R in the figure represents the image residual module in the middle.

3.1. PRNU and residual image training

The framework for implementing source camera identification (SCI) based on photo response non-uniformity (PRNU) is depicted in Fig. 3. PRNU is a fundamental feature in the literature of SCI, characterized by inherent properties resulting from hardware imperfections, and it remains consistent across images captured by the same device. Specifically, PRNU arises from variations in the thickness of silicon wafers across individual photosensitive elements within the sensor. Consequently, each pixel's performance exhibits slight deviations when exposed to identical lighting conditions. Notably, PRNU is widely acknowledged as a prominent image feature (fingerprint) in the literature, with numerous studies centered around its utilization.

The main difference between PRNU and other noise is that PRNU is multiplicative noise rather than additive noise. It has a strong correlation with image content, whereas additive noise is independently introduced. Therefore, by analyzing the principles of various noises, precise modeling of the image can be performed, enabling the design of feasible schemes for PRNU extraction. At the same time, to effectively estimate PRNU, the image model should not be overly complex. For a given image J taken by device d, each pixel is indexed by the spatial position $(x, y) \in \Omega$ and modeled as

$$J(x, y) = I(x, y) + I(x, y)K_d(x, y) + N(x, y) \tag{1}$$

where J represents the digital image obtained through the imaging device's capture and processing, $(x, y) \in \Omega$ denotes the spatial coordinates of each pixel. I corresponds to the original scene content captured by the imaging device. K_d represents the PRNU multiplicative noise component inherent to imaging device d, while N encompasses all additive noise components (including shot noise, quantization noise, Gaussian noise, etc.) that are statistically independent of K_d. As the unique fingerprint identifier in this digital image model for device authentication, K_d is simplistically modeled as a zero-mean multiplicative noise component independent of the image content I.

To achieve more accurate estimation of PRNU, it is necessary to eliminate interference from dominant image content. This is accomplished by processing the digital image J through a denoising filter to obtain a cleaner and clearer representation of the image content. Subsequently, the residual image R is obtained by subtracting the denoised image from the original image, thereby effectively preserving the K_d noise component. This procedure aims to enable more precise analysis of PRNU characteristics in the image while eliminating the influence of other interfering factors.

In order to derive the PRNU component K_d, the residual image R is obtained by passing image J through the de-noising filter F_d (as depicted in Fig. 3), for any $(x, y) \in \Omega$ we have

$$R(x, y) = J(x, y) - F_d[J(x, y)] \tag{2}$$

By Eqs. (1) and (2), the residual image R is accordingly

$$R(x, y) = I(x, y)K_d(x, y) + N(x, y) \tag{3}$$

In order to better estimate K_d, multiple images (say, M images for any device d) are usually used to obtain a reference PRNU (R_{PRNU}) and its maximum likelihood estimator is:

$$R_{PRNU} \approx K_d = \frac{\sum_{i=1}^{M} R_i J_i}{\sum_{i=1}^{M} J_i^2} \tag{4}$$

where, J_i represents the ith image, and R_i represents the residual image of the i-the image.

With R_{PRNU} in hands, we need to determine whether the test image t is taken by device d by calculating the cross-correlation function between the residual image R_t of t and the R_{PRNU} of device d.

$$\rho(K_d, R_t) = \frac{\sigma_{K_d R_t}}{\sigma_{K_d}\sigma_{R_t}} = \frac{\langle (R_t - \overline{R_t}), \left(K_d - \overline{K_d}\right)\rangle}{\|R_t - \overline{R_t}\|\|K_d - \overline{K_d}\|} \tag{5}$$

where, $\sigma_{K_d R_t}$ is the covariance of K_d and R_t, and σ_{K_d} and σ_{R_t} are the standard deviations of K_d and R_t, respectively. A larger $\rho(K_d, R_t)$ is desired as that implies a stronger correlation between R_t and K_d.

The PRNU of each device is unique and indelible since it is induced by hardware imperfections. Accordingly, we select PRNU features to guide model extraction and learn relevant features for source identification. So we propose a two-branch structure.

As PRNU is mainly in the high-frequency domain of the image, we apply a high-pass filter in one of the branches, which mainly focuses on the PRNU feature domain (see Fig. 2). The model **Siameformer** takes as input the image J and get:

$$\begin{aligned} &M_T(J(x, y)) = \\ &M_S(J(x, y)) + M_S(F_h[J(x, y)]), \forall (x, y) \in \Omega \end{aligned} \tag{6}$$

where M_T denotes the two-branch model, M_S denotes the single-branch model, and F_h denotes the high-pass filter.

Finally, the outputs from both branches are fused and processed to form a discriminative feature set for classification, as detailed in Section 3.3. It is worth noting that the proposed dual-branch architecture offers an inherent advantage in flexibility. While we employ a standard high-pass filter to emulate the initial step of focusing on high-frequency components (akin to traditional PRNU residual extraction), our framework is not constrained by this specific choice. The filter in the dedicated branch can be adapted or replaced (e.g., with a learnable filter bank or a more advanced frequency-domain decomposition module) to better isolate device-specific traces from complex image content in future work, potentially enhancing robustness against interference from strong edges and textures.

3.2. Global information and local features

Due to interference during network transmission, the PRNU feature domain may be obscured or lost. Therefore, when dealing with network images affected by the complex interference environment of social media, relying solely on PRNU and its related feature domain sets to guide model learning and forensics is far from sufficient. To address the complex interference situations in network images, deep models need to introduce more discriminative image fingerprint features to collaboratively accomplish device source forensics. Meanwhile, for image and feature domains in network images that are subject to interference or information loss, appropriate discarding is necessary to concentrate more computational resources on extracting and learning features from regions and feature domains where information is well-preserved. Therefore, this method introduces a multi-head attention mechanism, which simulates the complete process of depthwise separable convolution through self-attention, thereby integrating CNN and ViT, as well as combining local features and global information. Ultimately, the global information of network-interfered images is used to guide the model in extracting and learning local fingerprint features for device source forensics.

The multi-head self-attention mechanism decomposes each vector in the input feature sequence and assigns them to individual attention heads. This allows each attention head to compute a set of attention weights and learn its own representation of global feature information. The representation results from all attention heads are then concatenated to produce the final global feature extraction and learning outcome of the multi-head self-attention module. By employing the multi-head self-attention mechanism, the expressive power and generalization performance of the deep model can be effectively enhanced, thereby improving the source forensics performance of the model. Its structural framework is illustrated in Fig. 4. The module primarily consists of Dynamic Position Embedding (DPE), Multi-Head Attention (MHA), and Feedforward Neural Network (FNN).

To enable the deep network model to better learn the input image fingerprint features, making features in different regions more distinct and accurate, thereby improving the performance and effectiveness of self-attention mechanisms in extracting global features, both the local and global feature attention modules incorporate convolutional layers of varying sizes to map features into deeper dimensions. This facilitates the deep model in recognizing and leveraging more fingerprint features across different scales. This enhancement allows the deep model to better distinguish different image fingerprint information and enhance its robustness. After obtaining the mapped input feature vector X_{in} through different convolutional layers, dynamic positional encoding is applied to the vector:

$$\mathrm{DPE}(X_{in}) = \sum_{p=0}^{k-1}\sum_{q=0}^{k-1} R(X_{\mathrm{in}})_{p-\frac{k-1}{2}, q-\frac{k-1}{2}} \cdot K_{k-1-p, k-1-q} \tag{7}$$

where $R(\cdot)$ extends the tensor X_{in} to match the dimensions of the original image, i.e. $H \times W \times C$. Following this, a two-dimensional deep

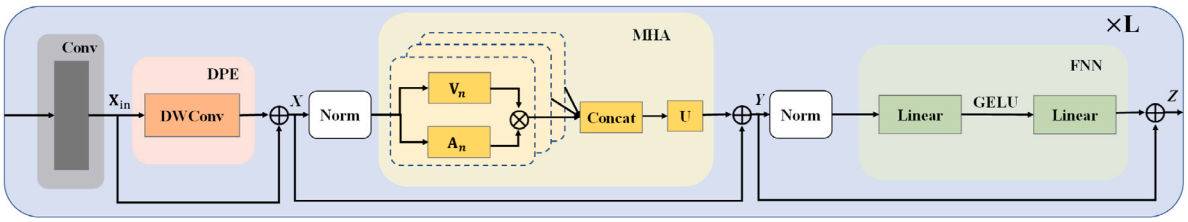


Fig. 4. Local/global feature attention module details.

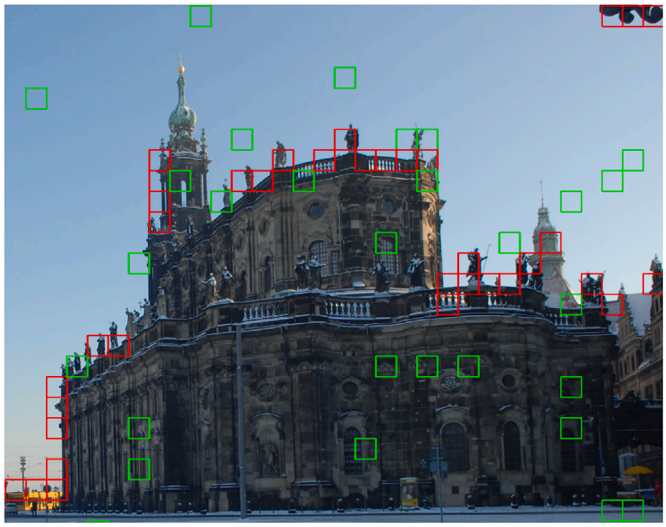


Fig. 5. Image content-feature correlation visualization results. (The red and green boxes focus on different image content and features, respectively. The red box focuses on edge and text content, and the green box focuses on semantic content.)

convolution is performed using a convolution kernel of size $k \times k$ and zero padding of $(k-1)/2$. The matrix K represents the positional encoding matrix, where $K_{k-1-p,k-1-q}$ denotes an element in the positional encoding matrix that corresponds to the reverse relative position with respect to the sub-matrix of input data.

The dynamic positional encoding module based on depthwise convolution enhances the model's perception of spatial information within features, while also offering greater flexibility and applicability to input feature vectors of any size. Additionally, the zero-padding incorporated in this module progressively queries boundaries to obtain their absolute positions. This dynamic encoding mechanism helps improve the representational capacity and generalization ability of the deep model. In the next step, the positional encoding matrix $\mathrm{DPE}(X_{in})$ output by the DPE module and the input feature mapping vector X_{in} are summed and then fed into the multi-head attention module:

$$Y = \mathrm{MHA}(\mathrm{Norm}(X)) + X \tag{8}$$

$$\mathrm{MHA}(X_0) = \mathrm{Concat}(R_1(X_0); \cdots; R_N(X_0)) \cdot U \tag{9}$$

$$R_n(X_0) = A_n(X_0) \cdot V_n(X_0), \text{ for } n \in 1, 2, \ldots, N \tag{10}$$

where $X = \mathrm{DPE}(X_{in}) + X_{in}$, $X_0 = \mathrm{Norm}(X)$. $R_n(\cdot)$ denotes the nth attention head. U is the weight parameter matrix of magnitude $C \times C$, integral over N heads. $\mathrm{Norm}(\cdot)$ denotes the batch normalization operation. For different objectives and functions of the attention module for local and global features, we make different modifications to R_n to focus on different regions.

The improvement approach for the local feature attention module is primarily based on the correlation between image content and features [38]. Digital image content contains rich semantic, color, edge, texture and other feature information. For example, in Fig. 5 the red and green boxes focus on different image content and features, respectively. The red box focuses on edge and text content, and the green box focuses on semantic content. Choosing different image block input will bring different performance to the model. This correlation also exists for fingerprint feature PRNU [39]. The emergence and development of ViT and its self-attention mechanism make it possible to capture this content-feature correlation. In the experiment part, the ablation experiment of the ViT branch of **Siameformer** was also carried out to prove the differences and reasons of source recognition results caused by the use of ViT and its self-attention mechanism.

In the local feature attention module, we apply spatial attention in MHA to learn this content-feature correlation. The attention operation A_n can be expressed as follows:

$$A_n^{local}(X_i, X_j) = softmax[\frac{Q_n(X_i)\left(K_n\left(X_j\right)\right)^T}{\sqrt{D_h}}] \tag{11}$$

where $Q_n(X_i) = X_i W_n^Q$ and $K_n(X_i) = X_j \cdot W_n^K$ respectively represent the query and key vectors of feature vector subsequence X_i obtained by multiplying X with query projection matrix and key query matrix, $\sqrt{D_h}$ represents the dimensional size of the nth attention head. To emphasize the correlation between content and features, rather than content relationships, and to reduce redundant calculations, we constrain each attention head to focus on a small windowed region for extracting local features across spatial locations. This approach is akin to the convolution operation in Convolutional Neural Networks.

In the case of spatial self-attention implemented within a small window, the linear transformation $V_n(X) = XW_n^V$ computes the value vector of X, which can be viewed as a pointwise convolution (PWConv). The expression for $A_n^{local}(X_i, X_j)$ is as follows:

$$A_n^{local}(X_i, X_j) = a_n^{i-j} \tag{12}$$

where the parameter matrix a_n represents the learnable weights, and $i-j$ denotes the relative positions of tokens i and j. Notably, A_n^{local} can be considered as the weight parameter matrix of each head operation $V_n(X)$. In this context, $R_n(X) = A_n^{\mathrm{local}}(X) \cdot V_n(X)$ effectively functions as a deep convolution. The combination of all attention heads through Eq. (9) is equivalent to a point convolution. This pattern bears a resemblance to the pw-dw-pw convolution pattern used in MobileNet.

Hence, the local feature attention module can be interpreted as a MobileNet block to some extent. By simulating a CNN with multi-head spatial self-attention and implementing convolutional operations, the model can efficiently leverage the content-feature correlation in the spatial dimension, effectively extracting features for the SCI task.

When the **Siameformer** deep model has acquired the capability to effectively extract local features, the next critical challenge for this method becomes how to identify the remaining local feature information in network images. To address this issue, **Siameformer** fully leverages the key advantage offered by the self-attention mechanism: global information modeling. Therefore, in the global feature attention module, channel attention is employed for global feature extraction.

In channel attention, each channel subsequence of the input vector contains holistic feature information. On the one hand, when computing correlations between different channels, the attention heads

naturally account for the influence of spatial positions within each channel, thereby capturing global information. On the other hand, the attention heads can assess the importance of information across channels, enhancing significant feature channels while suppressing irrelevant ones. This process better preserves the effective fingerprint information within the feature vectors, removes noise and redundant information, and more efficiently captures global information. Thus, channel attention enables superior global information capture and enhances the accuracy and generalization capability of the deep model.

Spatial attention primarily learns by gathering information across spatial dimensions within patches, whereas channel attention learns by gathering information along the channel dimension. Given this, we can conduct attention learning after patch transposition, where each transposed patch can be regarded as a representation of global information. As a result, the formulation of the attention module based on global features can be derived as follows:

$$A_n^{global}(X_i, X_j) = softmax[\frac{(Q_n(X_i))^T K_n(X_j)}{\sqrt{C_g}}] \tag{13}$$

where C_g denotes the size of the channel feature dimension.

The final step of the local/global feature attention module involves further feature mining and processing using a feed-forward neural network. This step is crucial for enhancing the model's expressiveness and prediction performance.

$$Z = \mathrm{FNN}(Norm(Y)) + Y \tag{14}$$

Ultimately, to achieve local feature extraction and leverage global information guidance, we combine CNN and ViT by iteratively stacking the local feature attention module and global feature attention module. This interplay between local features and global information serves as a crucial guide for model learning, enhancing its robustness for the task at hand.

3.3. Network weight sharing

The weight-sharing mechanism was initially introduced in the convolution operation of CNNs. Its purpose was to address the challenge of managing a large number of weight parameters for neurons, making adjustments and transfers difficult. By sharing certain weight values among neurons, it becomes feasible to compute the same weight parameter (finite convolution kernel) for different positions of the feature map.

Another application of weight sharing emerged with the advent of Siamese networks [40]. Siamese networks are commonly used to assess the similarity between two input sequences and have been utilized in various domains, such as face identification and semantic matching. The underlying concept involves mapping two data sequences into the same space through shared parameters, and then calculating the similarity between the two mapped quantities. The key objective is to minimize the distance between similar samples while maximizing the distance between different samples.

Drawing inspiration from the weight-sharing idea in Siamese networks for comparing sample similarity, we introduce this concept to the SCI domain. Our experiments in Section 4.2 confirm the feasibility of this approach. Leveraging the powerful camera fingerprint feature PRNU, we prioritize this fingerprint feature by sharing weights. In Fig. 2, the solid blue boxes represent the two network branches sharing weights in **Siameformer**. By calculating the similarity of features from these branches, they gradually approximate to the same, ultimately yielding PRNU features after integrating local and global information. These features are then fed into the classifier for classification, implementing the residual image training strategy from a different perspective.

The similarity comparison based on the weight-sharing mechanism is primarily achieved through a contrastive loss function, and the shared weight parameters in the branch networks are updated via gradient backpropagation. Let the input digital image pair to the model be J_1 and $J_2 = F_h[J_1]$. After mapping, the image feature vector pair fed into the network branches is X_1 and X_2. The contrastive loss functions (L_{Ct}) are defined as follows:

$$L_{Ct}(X_1, X_2) = \sum_{n=1}^{N} Y D_W^2 + (1-Y)max(m - D_W, 0)^2 \tag{15}$$

$$D_W(X_1, X_2) = \|X_1 - X_2\|_2 = (\sum_{i=1}^{P}(X_1^i - X_2^i)^2)^{\frac{1}{2}} \tag{16}$$

where, D_W represents the Euclidean distance between an image feature pair, as expressed in Eq. (16). Here, N stands for the number of sample pairs, P denotes the feature dimension size, and Y is a binary label indicating whether X_1 and X_2 were taken by the same camera. When $Y = 1$, it indicates that the image pairs are captured by the same camera, while $Y = 0$ suggests otherwise. The parameter m represents the preset threshold value. As a result, the loss function L_{Ct} can be simplified in this scenario as follows:

$$L_{Ct}(X_1, X_2) = \sum_{n=1}^{N} Y D_W^2 \tag{17}$$

When the Euclidean distance between pairs of image feature vectors gradually decreases, the contrastive loss function L_{Ct} also gradually decreases. When the Euclidean distance reaches its minimum achievable value, it indicates that the similarity between similar feature vector pairs has reached its highest point, and the value of the contrastive loss function is minimized, indicating that the model no longer requires further optimization. At this point, the deep model **Siameformer** achieves the integration of local features and global information in the relevant feature domain of PRNU fingerprints. The obtained PRNU features are then fed into a classifier and classified based on the cross-entropy loss function:

$$L_{Cr} = -\sum_{k=1}^{K} y_k log(p_k) \tag{18}$$

where K represents the number of classification classes, and y and p denote the real labels and predictive labels, respectively. By using the cross-entropy loss function, the penalty for incorrect predictions in the sample classification by the deep model is minimized. Then, by combining equations Eqs. (17) and (18), the final loss function can be obtained.:

$$L = L_{Ct} + L_{Cr} \tag{19}$$

Finally, combining Eq. (6), the process of training the residual graph for the digital image implementation of the **Siameformer** input-depth model can be derived:

$$M_T(J(x,y)) \approx M_S(R(x,y)), \forall (x,y) \in \Omega \tag{20}$$

Therefore, for the original or network-interfered image J input into the deep model **Siameformer**, this method incorporates a high-pass filter to enhance the model's focus on features within the PRNU-related fingerprint domain. Ultimately, through the weight-sharing mechanism, the deep model receives feature vector inputs akin to residual maps. This guides **Siameformer** to better extract local features and integrate global information for digital image source forensics.

4. Experiments and analysis

In order to evaluate the source identification performance and robustness of the model **Siameformer** proposed in Section 3, we conducted a large number of experiments on Dresden [41], Vision [42] and Forchheim [43] datasets. These three datasets are representative of the SCI domain. The detailed experimental results are given below. Detailed experimental results under the three datasets, referred to as confusion matrices, can be found in Appendix.

4.1. Implementation details

Datasets: (1) The Dresden dataset consists of 17,000 images taken by 74 digital cameras from 27 different models. (2) With the popularity of smartphones, the Vision dataset consists of 34,427 images taken by 35 smartphones of 11 brands and their social versions. (3) For the problem of social network interference images, the Forchheim dataset consists of 23,000 images captured by 27 different models and transmitted through social networks.

Comparison with existing methods:

(1) Camera model identification with Residual Neural Network (CRNN) [32];
(2) Deep learning for source camera identification on mobile devices (DSM) [36];
(3) Efficient Source Camera Identification with Diversity-Enhanced Patch Selection and Deep Residual Prediction (ESN) [38];
(4) Adversarial Analysis for Source Camera Identification (AAS) [37];
(5) Blind Source Camera Identification of Online Social Network Images Using Adaptive Thresholding Technique (BSAT) [44];
(6) Multiscale Content-Independent Feature Fusion Network for Source Camera Identification (MCFN) [45];
(7) A robust PRNU-based source camera attribution with convolutional neural networks (RPS) [46];
(8) CNN-Based Fast Source Device Identification (PCN) [47];
(9) RemNet: Remnant Convolutional Neural Network for Camera Model Identification (RN) [48];
(10) Enhanced Source Camera Identification Using Dual Pathway Processing and Spatial Attention Module (DPPASM) [49];
(11) EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks (EN) [50].

It should be noted that since SOTA methods are not open source and cannot be copied, some SOTA methods do not have test results on the second and third datasets, so there is no data on the second and third datasets.

Experiment evaluation metrics: In this paper, the accuracy rate of source forensics, confusion matrix, Floating Point Operations (FLOPs) and the number of parameters are used as evaluation indicators to evaluate the performance of different method models. Among them, the source forensics accuracy is the performance indicator of the evaluation method in correctly identifying the source of the image. It usually represents the ratio of the number of correctly identified images to the total number of samples. The confusion matrix is a table used to further evaluate the forensics performance of a method source, where the rows represent the actual category, the columns represent the predicted category, and each cell of the matrix represents the probability that the actual category is identified as the predicted category. FLOPs and parameter count are mainly used to measure the model complexity. In the experimental results, for the highest and second-highest source forensics accuracy, this section uses bold or underlined in the table.

Experiment details: In this paper, the same experimental Settings as the existing source forensics methods [37] are used to conduct the digital image source forensics experiment to ensure the fairness of the comparison. During the process of device source forensics experiments on images, for each model/individual category of digital image samples, the training images and test images account for 70% and 30% of all images in that category, respectively. To be consistent with all methods, each category of digital images is cropped to a small size of 256×256 before being input into the model for training or testing. All the experiments in this article were conducted under the Tensorflow framework using the NVIDIA GTX 2080Ti GPUs. During the training process, the loss functions of the two methods proposed in this paper were optimized through the Adamw algorithm. The learning rates were both set to 0.0005, the Batch training Size was set to 24, and a total of 200 rounds of training were conducted.

Table 1
Individual camera source identification comparison experiment results of original scenes on Dresden dataset (%).

Method	Our	AAS	DSM	CRNN	MCFN
ACC	**98.21**	96.13	91.60	94.30	94.24

In this section, a total of 6 groups of comparative experiments were carried out.

(1) In the original scene, the performance comparison experiment is carried out on the individual source forensics of digital images.

(2) In the original scene, the performance comparison experiment is carried out on the same type of digital image individual source forensics.

(3) In the original scene, the performance comparison experiment is carried out on the model source forensics of digital images.

(4)Simulation and comparison experiments of interference modes in real scenes are set up to verify the robustness of the proposed method. In this experiment, a total of 6 common interference modes in real scenes were selected to set up the experimental group. Specifically, the six interference modes are: JPEG compression (quality factor 80), noise disturbance (Gaussian noise), random cropping (512×512 cropping followed by 224×224 center cropping to get a small size image inconsistent with the edge information of the training set, and select and lose part of it according to the texture and semantic information. This makes the size and edge texture information of the cropped image different from the training image) and the mixed interference of high and low quality (that is, the network transmission through the two transmission methods of Facebook). The digital images in each experimental group were disturbed by one of the interference modes.

(5) In order to verify that the two methods proposed in this paper can solve the problems existing in network image source forensics, a comparative experiment of source forensics in a real scene is set up. A total of five major social networks (Facebook, Instagram, Telegram, Twitter and WhatsApp) were selected for the experiment.

(6) In order to prove the effectiveness of the lightweight strategy adopted in the proposed method, the method is compared with the existing one The computational complexity and the number of parameters are compared with the large ViT model.

Finally, in addition to the comparison experiments, the ablation experiments of each module and some parameters of the two methods are also carried out in this section.

4.2. Experimental results and analysis

To evaluate the performance of the proposed **Siameformer**, this study conducted fair comparisons with other existing methods under the same dataset and experimental settings. A diverse set of individual digital image categories was selected from three image datasets for individual device source forensics and camera model source forensics as comprehensively as possible. Various methods were evaluated under this experimental setup to assess their accuracy objectively.

As shown in Table 1, the performance comparison results of individual device source identification in the original scenario for the Dresden image dataset between the two methods proposed in this paper and the four digital image source forensics methods (AAS, DSM, CRNN, MCFN) are presented. Among them, **Siameformer** achieved the highest source identification accuracy at 98.21%.This demonstrates that integrating traditional machine learning approaches based on PRNU to construct residual images with deep learning is feasible and effective. Unlike other fingerprint features, PRNU features are unique, allowing the model to more effectively distinguish between different device identities. As the **Siameformer** model deepens, there is also increasing focus on the PRNU feature domain, thereby achieving more effective individual source forensics.

Table 2
Individual smartphone source identification comparison experiment results of original scenes on Vision dataset(%).

Method	Our	AAS	DSM	ESN	BSAT	MCFN	RPS	DPPASM
ACC	**95.45**	93.83	87.23	84.34	95.14	92.35	93.48	91.41

Table 3
SCI comparison experiment results of original scenes on Forchheim dataset (%).

Method	Our	AAS	DSM	RN	EN
ACC	**97.08**	95.05	83.66	93.80	96.30

With the development and popularity of smartphones, it is not enough to identify the source device of a camera image. As shown in Table 2, the performance comparison results of **Siameformer** and seven SCI methods including AAS, DSM, ESN, BSAT, MCFN, RPS and DPPASM in the Vision dataset are shown. It can be seen that in this dataset, the source identification ACC of **Siameformer** is 95.45%. As can be seen from the results, **Siameformer** can also achieve superior performance for images taken by mobile phones.

Similar to the Vision dataset, the Forchheim dataset is also designed for mobile image source identification. At the same time, it is more focused on the social version of mobile images. Table 3 shows the performance comparison results of **Siameformer** and the four SCI methods, AAS, DSM, RN and EN, on the Forchheim dataset. It can be seen that **Siameformer** achieves 97.08% ACC for source identification under this dataset.

Combining the performance of **Siameformer** on the three datasets, we further analyze the internal reasons for its such distribution. The source identification ACC of **Siameformer** on the Dresden, Vision and Forchheim datasets is 98.21%, 95.45% and 97.08%, respectively. By comparison, it is found that **Siameformer** has different levels of performance degradation on the latter two datasets. There are three possible reasons for this analysis:

(1) Vision and Forchheim datasets are sample datasets of mobile phone images. Compared to cameras, mobile phone lenses have lower optical performance and smaller sensor sizes. So the phone captures less light and less detail. It can be conjectured that PRNU may be more ambiguous and difficult to extract for mobile phone image samples. At the same time,it is worth discussing whether the sample of mobile phone images contains other features that camera images do not.
(2) Since the number of categories in Vision dataset is more than that in the other two datasets, the increase in the number of categories of image models increases the burden on the classifier, resulting in a decline in the accuracy of source identification.
(3) Since the training samples of each category in Forchheim dataset are less than the other two datasets, the reduction of training samples has a negative impact on the classifier, resulting in a decline in the accuracy of source identification.

To thoroughly evaluate the performance of **Siameformer** in individual source identification and address the performance degradation issue when identifying sources from the same model in individual source identification, this section designs comparative experiments under the original scenario for individual source identification of the same model. The experiment selects different individuals of the same digital camera or smartphone model from three image datasets for individual source identification. In this context, individual source identification for the same model can be viewed as a binary or ternary classification problem. According to Table 4, **Siameformer** achieves the highest source identification accuracy for individual source identification of the same model in the original scenario. It shows an improvement of 9.44%, and 22.42% over AAS, and DSM, respectively. This indicates that **Siameformer**, primarily driven by PRNU fingerprint features, can effectively distinguish digital images generated by different individuals of the same model in individual source identification tasks.

Table 4
Comparative experimental individual level source identification by model level SCI method (%).

Method	Our	AAS	DSM
ACC	**88.72**	78.28	65.30

Table 5
Comparative experimental individual level source identification by model level SCI method (%).

Method	Dresden	Vision	Forchheim	Average
AAS	96.27	94.04	95.48	95.26
DSM	91.78	88.52	84.98	88.43
Our	**98.37**	**95.53**	**96.71**	**96.87**

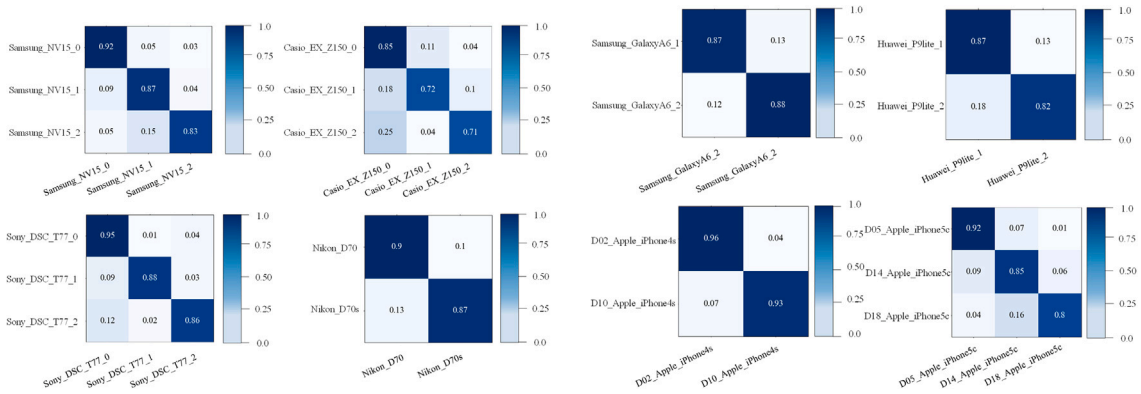
Since **Siameformer** is implemented to some extent based on PRNU noise, all existing PRNU based methods are aimed at source camera individual identification. In order to evaluate whether the model SCI method **Siameformer** based on PRNU has the ability of individual identification, it also aims to solve the problem of source identification level increase in model identification. We design an individual-level source identification task, and extracted the images of camera/phone model individuals with increasing source identification level in the three datasets for source identification. In this case, the individual-level source identification task can also be regarded as a binary or triple classification problem. As shown in Table 4, among the SCI algorithms oriented to model level, the individual source identification ACC of **Siameformer** is 9.44% and 22.42% higher than that of AAS and DSM. It is shown that **Siameformer** can still efficiently classify individuals of the same type in individual identification.

Considering the potential differences between camera and smartphone images and the influence of the number of individual devices, this experiment selected four classes of camera models (Samsung_NV15, Casio_EX_Z150, Nikon_D70, and Sony_DSC_T77) and four classes of smartphone models (Samsung_GalaxyA6, Huawei_P9lite, Apple_iPhone4s, and Apple_iPhone5c) from three image datasets. Each model class contained 2 to 3 individual devices. As shown in Fig. 6 (a) and 6(b), the confusion matrix depicts the results of **Siameformer** in the experiment on individual source identification of the same model.

It can be observed that targeted learning for different individuals of the same model effectively reduces the misclassification rate of the model for different individuals within the same model type. **Siameformer** demonstrates consistent source identification accuracy for individual devices within the same model of cameras or smartphones. This validates that PRNU features are effective in distinguishing different individuals of the same model in cameras or smartphones.

In order to verify the performance of **Siameformer** in the original scenario of model source forensics, this section designed the experiment. As shown in Table 5, when considering only different device models but not different individuals of the same model, there is a difference of 2.84% between the highest and lowest accuracy of **Siameformer** in the forensics of model sources under the three image datasets. The algorithm framework based on self-attention mechanism combining CNN and ViT performs well and stably.

To evaluate the robustness of **Siameformer** against different interference patterns in real-world scenarios, this paper designed a simulation-based comparative experiment that simulated common interference patterns in digital images, including compression, noise, and cropping. The experiment included two examples of mixed interference: high-quality and low-quality transmission modes from Facebook and conducted simulations using the Vision dataset. It should be added that some methods (such as ESN, MCFN, etc.) are not open source and do not record experimental data through Gaussian noise or Random cropping, so the experimental results are not reported.



(a) Individual source identification of camera images.

(b) Individual source identification of smartphone images.

Fig. 6. Siameformer Confusion matrix of individual source forensics with the same model.

Table 6
Comparison of SCI experimental results in real distortion scenarios (%).

Method	Our	AAS	DSM	ESN	BSAT	MCFN	PCN
Original	**95.45**	93.83	87.23	84.34	95.14	92.35	91.41
JPEG 80	**92.50**	77.07	77.83	74.28	87.91	79.63	80.21
Gaussian noise	**93.18**	90.56	80.24	–	84.35	–	–
Random cropping	90.00	82.82	81.43	–	**90.14**	–	–
Facebook_HQ	**92.29**	74.68	70.45	67.16	80.26	70.26	71.81
Facebook_LQ	**89.77**	69.23	66.86	61.94	75.08	63.45	62.31

Table 7
Comparative experiments on source identification after social network propagation under the Forchheim dataset (%).

Method	Our	AAS	DSM	RN	EN
Original	**97.08**	95.00	83.66	93.80	96.30
Facebook	**92.46**	72.57	58.69	36.00	51.10
Instagram	**83.28**	65.46	47.33	48.90	67.50
Telegram	**82.42**	73.83	69.71	52.80	73.10
Twitter	**93.32**	75.17	69.76	74.20	93.20
WhatsApp	**92.79**	81.48	69.84	50.80	72.90

As shown in Table 6, we perform a total of 5 sets of robust source identification experiments in the distorted scenarios. In order to be consistent with the selected comparison method [36–38,44,45,47,51], we conduct experiments with the same experimental setup on the Vision dataset. **Siameformer** achieves the highest source identification accuracy in the original scenario, with performance only declining by 2.27% to 5.68% under interference, while other methods show a decline of 7.23% to 16.76%. In the random cropping group, **Siameformer's** source identification accuracy is slightly lower than BSAT. The analysis in this paper suggests that the main reason is that random cropping disrupts edge information and increases the difficulty of extracting global information. However, in terms of overall source identification accuracy, **Siameformer** still performs well in the simulation experiments and exhibits better robustness than other methods. The primary reason is that when images are interfered with and lose part of their information or features, **Siameformer** can utilize the local feature attention module to mine and learn the image fingerprint features that distinguish different devices. At the same time, it models the global information of the image through the global feature attention module, focusing more on image domains and feature domains that are unaffected or only slightly affected by interference, thereby enhancing its robustness.

The distortion of images transmitted through social networks is usually complex (a combination of multiple distortion operations). To this end, we set up a source identification experiment after social network transmission in the Forchheim dataset, and the specific results are shown in Table 7. After the transmission of five social media, Facebook, Instagram, Telegram, Twitter and WhatsApp, **Siameformer**'s source identification ACC decreased by 4.62%, 13.8%, 14.66%, 3.76% and 4.29%, respectively. Compared with the other four methods, ACC is reduced by up to 57.8% to at least 3.1%, **Siameformer** is still able to maintain stable source identification performance in the face of complex social network transmission environment and is robust. A horizontal comparison shows that Twitter and WhatsApp retain more image information during transmission, followed by Facebook, Instagram and Telegram.

Table 8
Experimental results of the influence of different filter on the performance of Siameformer (%).

Method	FLOPs	Parameters
CoAtNet-3	34.7G	168.0M
LV-ViT-L	59.0G	150.0M
Our	**27.2G**	**50.1M**

To evaluate the effectiveness of the lightweight strategy in our method, we conducted comparisons with two established hybrid architectures combining ViT and CNN (CoAtNet-3 [52] and LV-ViT-L [53]).

As shown in Table 8. It can be seen that the computational complexity and parameter number of **Siameformer** are significantly lower than the existing model combining ViT and CNN. In addition, combining CNN and ViT models by fusion module and cross-attention mechanism is much less costly than simulating convolution operation by self-attention mechanism.

More importantly, our approach of combining CNN and ViT through a lightweight fusion module and cross-attention mechanism proves to be far more computationally economical than strategies that simulate convolution operations purely via self-attention mechanisms. Our experiments and comparisons clearly show that the performance gain comes from our fusion strategy, rather than from simply increasing model capacity. Therefore, the primary focus of this paper is to validate

Table 9
Experimental results of the influence of siamese branches and filters on Siameformer (%).

Siamese-branches	High-pass filter	ACC
×	×	86.71
×	√	92.33
√	×	88.90
√	√	**98.21**

Table 10
Experimental results of the influence of different filter on the performance of Siameformer (%).

Filter	Dresden	Vision	Forchheim
Gaussian	91.03	91.75	93.96
High-pass	**98.21**	88.03	91.53
Bilateral	93.65	**95.45**	**97.08**
Sharpening	76.94	76.75	84.04
Low-pass	87.05	81.25	74.23

the effectiveness of this fusion mechanism against other hybrid methods, rather than on ablating individual components whose properties are already well established in prior work.

Next, we will discuss the results of the ablation experiments for **Siameformer**. As shown in Table 9, the impact of weight sharing in Siamese networks and the high-pass filtering based on PRNU feature selection on **Siameformer** are discussed. Under the combined influence of weight sharing and high-pass filtering, **Siameformer** demonstrates significantly improved performance. Compared to models without the Siamese shared branch and high-pass filter, **Siameformer** achieves an 11.5% increase in source identification accuracy. This confirms the effectiveness of the proposed algorithm that integrates global information guided by PRNU features with local features. When using only high-pass filtering, the source identification accuracy improves by 5.62%. High-pass filtering directs the model's focus more on the PRNU feature set, which is more discriminative in the high-frequency domain of images, effectively enhancing model performance. When extending the dual-model branch and incorporating weight sharing alone, the source identification accuracy improves by 2.19%. This is because using weight sharing alone is equivalent to training two identical models for classification. While this accelerates model convergence, it provides limited assistance in improving model performance.

As shown in Table 10, the impact of different filter types (across different feature domains) on **Siameformer** performance was studied across three image datasets. On the Dresden dataset, the **H-Siameformer** using a high-pass filter achieved the highest source identification accuracy at 98.21%. However, on the Vision and Forchheim datasets, the `H-Siameformer` did not perform as well, achieving only 88.03% and 91.53% accuracy, respectively. In contrast, the **Siameformer** using a bilateral filter (**B-Siameformer**) achieved the best performance on these two datasets, with accuracies of 95.45% and 97.08%, respectively.

This analysis reveals two key findings: First, the Dresden dataset (comprising camera images) exhibits more pronounced and easily extractable PRNU noise in the high-frequency domain. However, in smartphone images, high-frequency PRNU noise is obscured by edge texture information, resulting in blurred characteristics that limit performance improvements when models focus on this domain. Conversely, low-frequency PRNU noise becomes more distinguishable and extractable in smartphone images. Second, bilateral filtering demonstrates superior performance for smartphone image processing by simultaneously preserving edge information while considering both spatial and grayscale

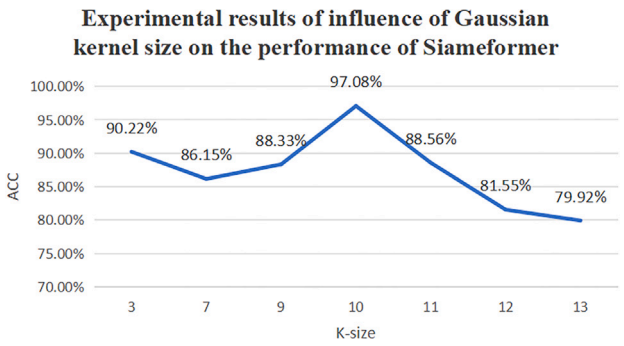


Fig. 7. Experimental results of influence of Gaussian kernel size on the performance of **Siameformer**.

domains.Comparative results indicate that the Gaussian-filtered **Siameformer** (**G-Siameformer**) also serves as a viable alternative. In contrast, models overemphasizing edge textures and image content (e.g., sharpened **Siameformer** and low-pass **Siameformer**) exhibit suboptimal performance due to their limited noise separation capabilities.

Finally, the impact of Gaussian kernel size (K-size) on model performance is experimentally investigated in the case where **B-Siameformer** achieved its best performance (referring to Fig. 7). Experiments are performed under the Dresden dataset. The size of the Gaussian kernel represents the standard deviation of the Gaussian distribution function in the spatial domain and is used to control the filtering effect. The **Siameformer** achieves the best performance with k-size = 11. When the Gaussian kernel size is gradually increased, the filter gradually enhances the texture details in the image, which interferes with the extraction of PRNU noise and leads to performance degradation. When the size of the Gaussian kernel is gradually reduced, the filter gradually enhances the edge information in the image and interferes with the PRNU noise extraction, leading to performance degradation. Of course, a certain amount of edge information is helpful for image source identification (K-size=7). Without interfering with PRNU noise extraction, both help to enrich the feature domain in the input network, thereby helping to improve performance (K-size=11).

5. Conclusion

Motivated by the lack of robustness of existing SCI methods and inspired by the superiority of self-attention mechanism in ViT, a robust SCI method based on CNN and ViT combined with social network is proposed. Experimental results show that the proposed method can achieve high-precision source camera identification in real-world distorting scenarios and effectively address the problem that the performance of the original image-based SCI method is somewhat degraded in the face of interference images. For the **Siameformer** proposed by us, not only the PRNU-like features can be applied to the model, but also more excellent features can be similarly transferred to achieve better performance. Therefore, classical features of PRNU-like networks are selected in this paper to justify the proposed holistic network framework. The highly robust source identification is achieved through the interaction between features and global information guidance, rather than through high-performance features. Based on the analysis of the contents described in the paper, there is still room for further improvement of our model. It includes: (1) using higher performance features to achieve the high performance phenotype of the method; (2) Explore the possibility of improving source identification through global information to improve robustness: combine CNN and ViT from the underlying principle or consider transfer learning; (3) Make more attempts in real interference scenarios: SCI in open Settings, SCI in large categories (500+) and SCI in small samples.

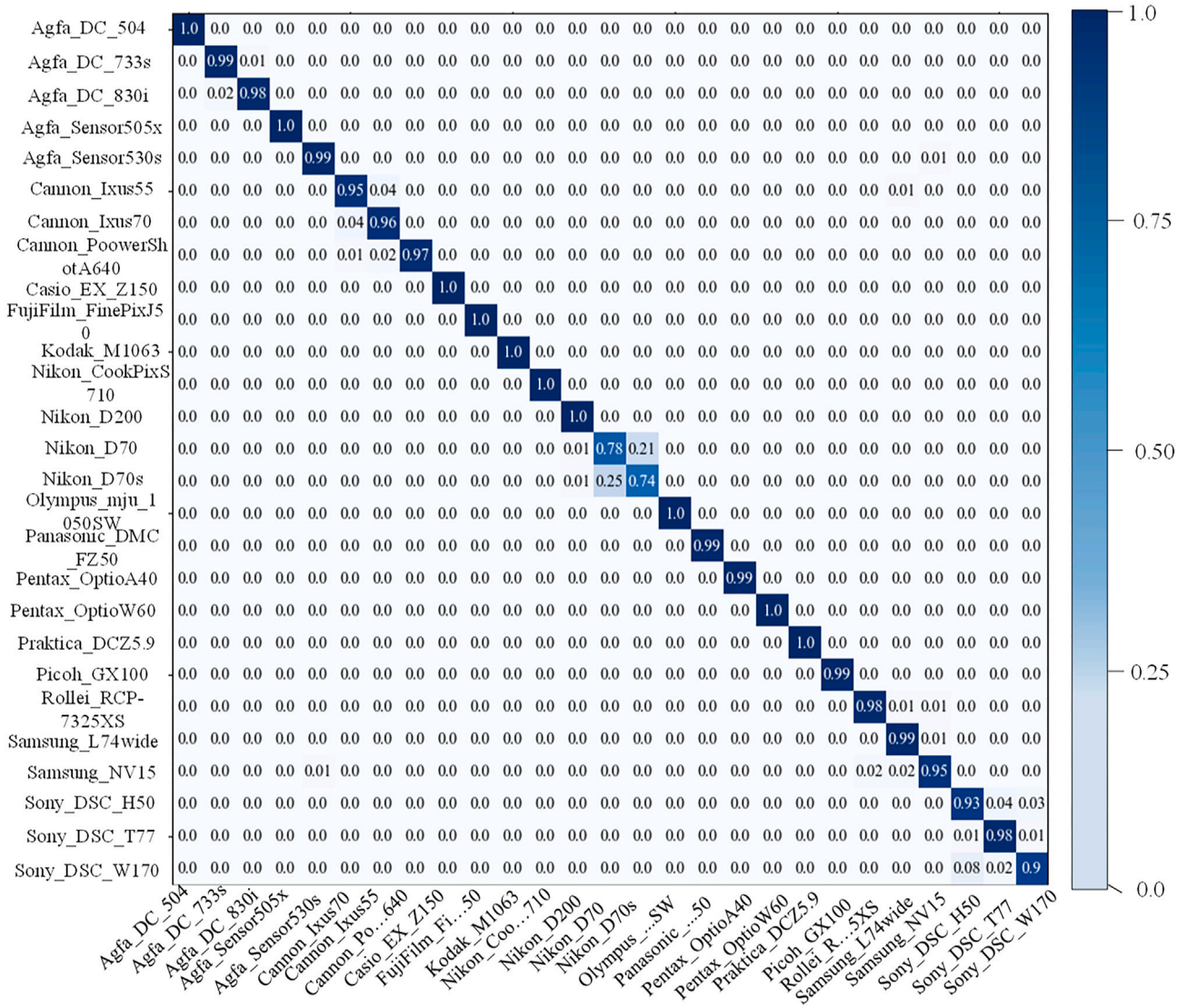


Fig. 8. Siameformer source identification confusion matrix on the Dresden dataset.

CRediT authorship contribution statement

Bo Wang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Jiaqi Chi:** Writing – review & editing, Visualization, Methodology, Formal analysis. **Weiming Zheng:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Fei Wei:** Writing – review & editing, Methodology, Conceptualization. **Yi Li:** Writing – review & editing, Supervision, Resources, Methodology, Conceptualization.

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.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62106037, No. 62076052), the Science and Technology Innovation Foundation of Dalian (No. 2021JJ12GX018), and the Application Fundamental Research Project of Liaoning Province (2022JH2/101300262).

Appendix. Confusion matrices

Here are detailed experimental results of Siameformer on three datasets called confusion matrices (see Figs. 8–10).

Data availability

The authors do not have permission to share data.

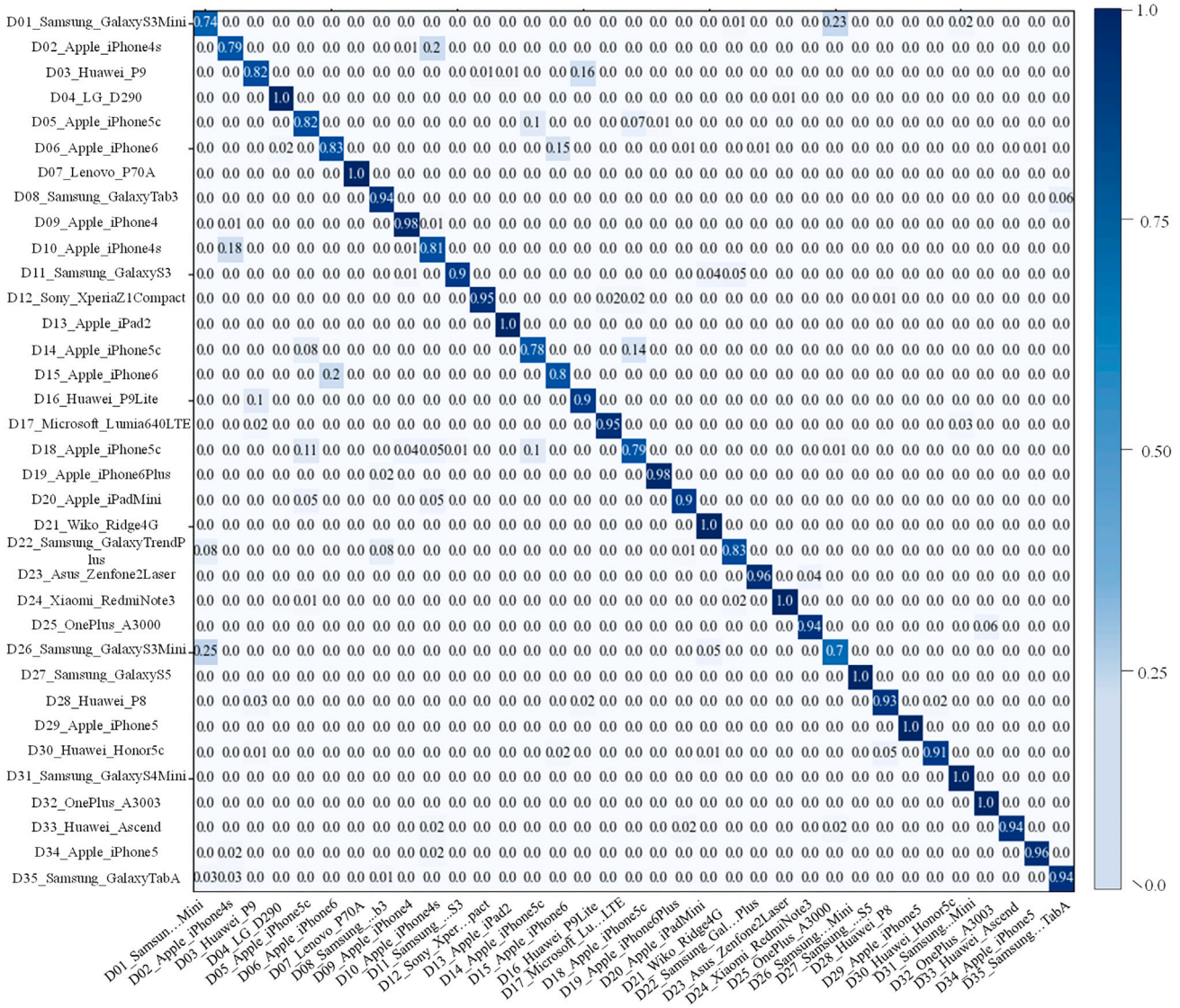


Fig. 9. Siameformer source identification confusion matrix on the Vision dataset.

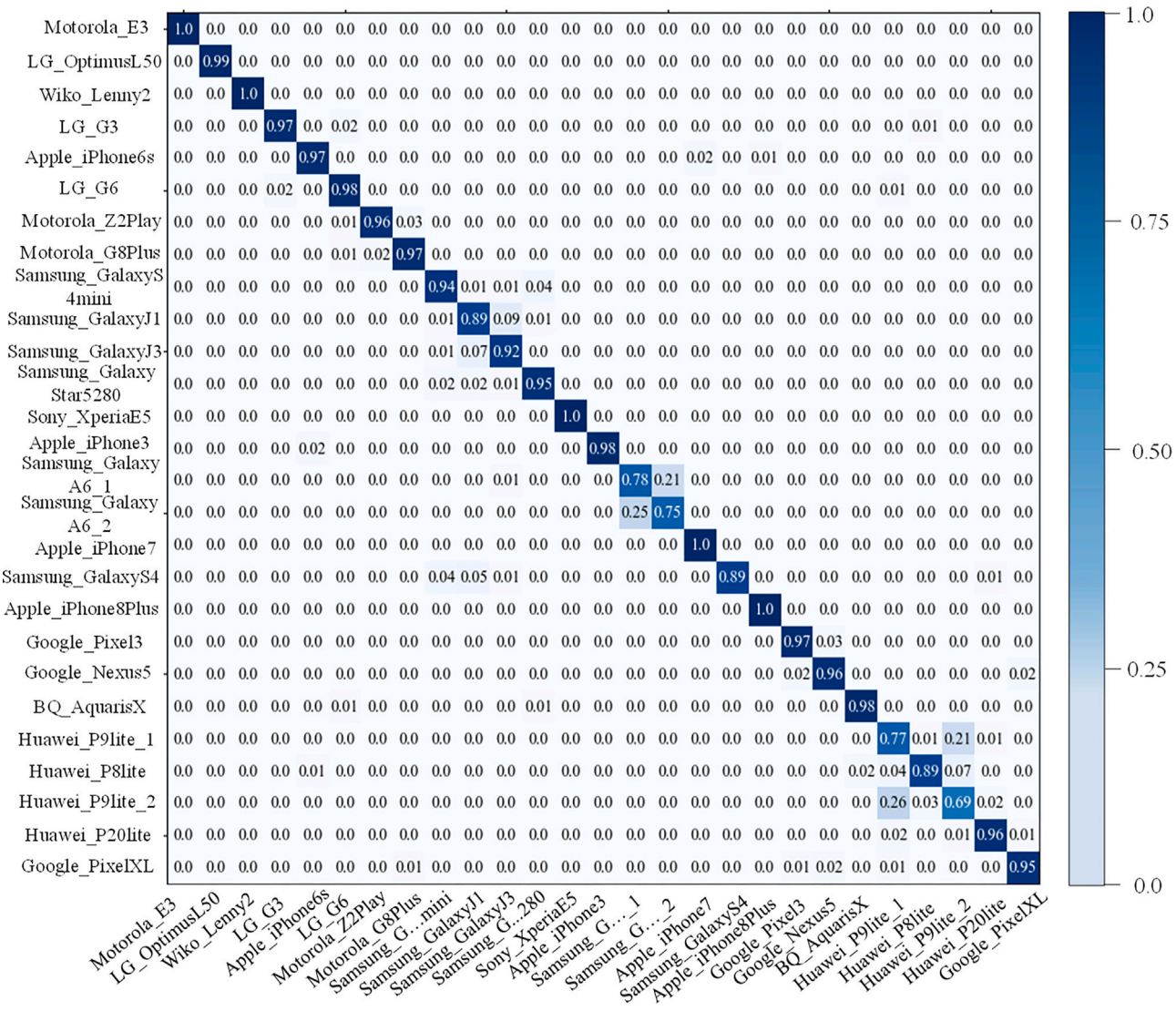


Fig. 10. Siameformer source identification confusion matrix on the Forchheim dataset.

References

[1] A. Agarwal, R. Keshari, M. Wadhwa, et al., Iris sensor identification in multi-camera environment, Inf. Fusion 45 (2019) 333–345.

[2] L. Baroffio, L. Bondi, P. Bestagini, et al., Camera identification with deep convolutional networks, Signal Process. Lett. 24 (3) (2016) 259–263.

[3] K. Rana, G. Singh, P. Goyal, MSRD-CNN: Multi-scale residual deep CNN for general-purpose image manipulation detection, IEEE Access 10 (2022) 41267–41275, http://dx.doi.org/10.1109/ACCESS.2022.3167714.

[4] M. Goljan, M. Chen, P. Comesana, et al., Effect of compression on sensor-fingerprint based camera identification, in: Media Watermarking, Security, and Forensics, 2016, pp. 1–10.

[5] C. Chen, M.C. Stamm, Robust camera model identification using demosaicing residual features, Multimedia Tools Appl. 80 (2021) 11365–11393.

[6] J. Bernacki, Robustness of digital camera identification with convolutional neural networks, Multimedia Tools Appl. 80 (19) (2021) 29657–29673.

[7] H. Lin, Y. Wo, Y. Wu, Robust source camera identification against adversarial attacks, Comput. Secur. 100 (2021) 079–102.

[8] G. Zhang, B. Wang, F. Wei, Source camera identification for re-compressed images: A model perspective based on tri-transfer learning, Comput. Secur. 100 (2021) 076–102.

[9] J. Lu, C. Li, X. Huang, C. Cui, M. Emam, Source camera identification algorithm based on multi-scale feature fusion, Comput. Mater. Contin. 80 (2) (2024).

[10] A.K. Jaiswal, R. Srivastava, Source camera identification using hybrid feature set and machine learning classifiers, in: Role of Data-Intensive Distributed Computing Systems in Designing Data Solutions, Springer, 2023, pp. 111–127.

[11] C. Sychandran, R. Shreelekshmi, SCCRNet: a framework for source camera identification on digital images, Neural Comput. Appl. 36 (3) (2024) 1167–1179.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[13] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, B. Guo, Cswin transformer: A general vision transformer backbone with cross-shaped windows, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12124–12134.

[14] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, Y. Qiao, Uniformer: Unifying convolution and self-attention for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 45 (10) (2023) 12581–12600.

[15] J. Lukas, J. Fridrich, M. Goljan, Digital camera identification from sensor pattern noise, Trans. Inf. Forensics Secur. 1 (2) (2006) 205–214.

[16] C. M., M. Goljan, J. Fridrich, Defending against fingerprint-copy attack in sensor-based camera identification, Trans. Inf. Forensics Secur. 6 (1) (2010) 227–236.

[17] X. Kang, Y. Li, Z. Qu, et al., Enhancing source camera identification performance with a camera reference phase sensor pattern noise, Trans. Inf. Forensics Secur. 7 (2) (2012) 393–402.

[18] V. Bruni, M. Tartaglione, D. Vitulano, Coherence of PRNU weighted estimations for improved source camera identification, Multimedia Tools Appl. 81 (16) (2022) 22653–22676.

[19] Y. Akbari, N. Almaadeed, S. Al-Maadeed, et al., PRNU-net: A deep learning approach for source camera model identification based on videos taken with smartphone[c], in: 2022 26th International Conference on Pattern Recognition, ICPR, IEEE, 2022, pp. 599–605.

[20] Manisha, C.T. Li, X. Lin, et al., Beyond PRNU: Learning robust device-specific fingerprint for source camera identification, Sensors 22 (20) (2022) 7871.

[21] B. Xu, X. Wang, X. Zhou, J. Xi, et al., Source camera identification from image texture features, Neurocomputing 207 (2016) 131–140.

[22] M. Kharrazi, H.T. Sencar, N. Memon, Blind source camera identification, in: Prof. of International Conference on Image Processing ICIP 04, vol. 1, IEEE, 2004, pp. 709–712.
[23] S. Bayram, H. Sencar, N. Memon, Source camera identification based on CFA interpolation, in: International Conference on Image Processing, IEEE, 2005, III–69.
[24] M.J. Tsai, C.S. Wang, J. Liu, Using decision fusion of feature selection in digital forensics for camera source model identification, Comput. Stand. Interfaces 34 (3) (2012) 292–304.
[25] Z. Deng, Gijsenij, J. Zhang, Source camera identification using auto-white balance approximation, in: 2011 International Conference on Computer Vision, 2011, pp. 57–64.
[26] K. San Choi, E.Y. Lam, K.K.Y. Wong, Automatic source camera identification using the intrinsic lens radial distortion, Opt. Express 14 (24) (2006) 11551–11565.
[27] G. Thomas, S. Pfennig, M. Kirchner, Unexpected artefacts in PRNU-based camera identification: a 'dresden image database' case-study, in: Proceedings of the on Multimedia and Security, 2012, pp. 109–114.
[28] F. Razzazi, A. Seyedabadi, A robust feature for single image camera identification using local binary patterns, in: IEEE International Symposium on Signal Processing and Information Technology, ISSPIT, IEEE, 2014, pp. 000462–000467.
[29] J. Xiao, Q. Dai, X. Xie, et al., Domain adaptive graph infomax via conditional adversarial networks, Trans. Netw. Sci. Eng. 10 (1) (2023) 35–52.
[30] A. Tuama, F. Comby, M. Chaumont, Camera model identification with the use of deep convolutional neural networks, in: 2016 IEEE International Workshop on Information Forensics and Security, WIFS, 2016, pp. 1–6.
[31] P. Yang, R. Ni, Y. Zhao, Source camera identification based on content-adaptive fusion network, Pattern Recognit. Lett. 12 (119) (2017) 125–127.
[32] Y. Chen, Y. Huang, X. Ding, Camera model identification with residual neural network, in: In Proc. IEEE Int. Conf. Image Process., ICIP, 2017, pp. 4337–4341.
[33] M. Rafi, Abdul, Kamal, Application of DenseNet in camera model identification and post-processing detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019, pp. 4337–4341.
[34] B. Bayar, M.C. Stamm, Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection, Trans. Inf. Forensics Secur. 4 (13) (2018) 2691–2706.
[35] I.J. Yu, D.G. Kim, J.S. Park, Identifying photorealistic computer graphics using convolutional neural networks, in: IEEE International Conference on Image Processing, IEEE, 2018, pp. 4337–4341.
[36] F. David, F. Narducci, S. Barra, et al., Deep learning for source camera identification on mobile devices, Pattern Recognit. Lett. 126 (2019) 86–91.
[37] B. Wang, M. Zhao, W. Wang, Adversarial analysis for source camera identification, Trans. Circuits Syst. Video Technol. 31 (11) (2021) 4174–4186.
[38] Y. Liu, Z. Zou, Y. Yang, et al., Efficient source camera identification with diversity-enhanced patch selection and deep residual prediction, Sensors 21 (14) (2021) 4701.
[39] E. Flor, R. Aygun, S. Mercan, et al., PRNU-based source camera identification for multimedia forensics, in: 2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science, IRI, IEEE, 2021, pp. 168–175.
[40] Z. Zhou, Y. Li, J. Li, et al., GAN-siamese network for cross-domain vehicle re-identification in intelligent transport systems, Trans. Netw. Sci. Eng. 10 (5) (2022) 2779–2790.
[41] G. Thomas, B. Rainer, The 'dresden image database' for benchmarking digital image forensics, in: Proceedings of the 2010 ACM Symposium on Applied Computing, 2010, pp. 1584–1590.
[42] D. Shullani, M. Fontani, M. Iuliani, et al., VISION: a video and image dataset for source identification, EURASIP J. Info. Secur. 15 (2017).
[43] B. Hadwiger, C. Riess, The forchheim image database for camera identification in the wild, in: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part VI, 2021, pp. 500–515.
[44] B.N. Sarkar, S. Barman, R. Naskar, Blind source camera identification of online social network images using adaptive thresholding technique, in: Proceedings of International Conference on Frontiers in Computing and Systems: COMSYS 2020, 2020, pp. 637–648.
[45] C. You, H. Zheng, Z. Guo, Multiscale content-independent feature fusion network for source camera identification, Appl. Sci. 11 (15) (2021) 6752.
[46] T. Nayerifard, H. Amintoosi, A. Ghaemi Bafghi, A robust PRNU-based source camera attribution with convolutional neural networks, J. Supercomput. 81 (1) (2025) 25.
[47] S. Mandelli, D. Cozzolino, P. Bestagini, Cnn-based fast source device identification, IEEE Signal Process. Lett. 27 (2020) 1285–1289.
[48] A.M. Rafi, T.I. Tonmoy, U. Kamal, RemNet: Remnant convolutional neural network for camera model identification, Neural Comput. Appl. 33 (2020) 1–16.
[49] A. Zaimen, A. Oulefki, F. Khelifi, T. Rabie, A. Bouridane, Enhanced source camera identification using dual pathway processing and spatial attention module, in: 2024 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, BDCAT, IEEE, 2024, pp. 198–203.
[50] M. Tan, Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, 2019, pp. 6105–6114.
[51] M. Kirchner, C. Johnson, Spn-cnn: Boosting sensorbased source camera attribution with deep learning, in: In 2019 IEEE International Workshop on Information Forensics and Security, WIFS, IEEE, 2019, pp. 1–6.
[52] Z. Dai, H. Liu, Q.V. Le, et al., CoAtNet: Marrying convolution and attention for all data sizes, in: Advances in Neural Information Processing Systems 34, NeurIPS 2021, Virtual, 2021, pp. 3965–3977.
[53] Z. Jiang, Q. Hou, L. Yuan, et al., All tokens matter: Token labeling for training better vision transformers, in: Advances in Neural Information Processing Systems 34, NeurIPS 2021, Virtual, 2021, pp. 18590–18602.