

RESEARCH

Open Access



Shoeprint retrieval network based on composite attention mechanism

Liman Liu¹, Yifei Han¹, Jinshan Tian¹, Han Wu², En Yu^{2*} , Qiling Tang¹ and Wenbing Tao²

*Correspondence:

yuen@hust.edu.cn

²National Key Laboratory of Science and Technology on Multi-spectral Information Processing, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, Hubei, 430074, China
Full list of author information is available at the end of the article

Abstract

Shoeprint retrieval is an important component of footprint identification, where it searches for shoe information with the same footprint texture in a shoe database based on the footprints at the crime scene. In most cases, criminals do not leave their shoes at the crime scene, which means that a shoeprint retrieval is necessary to identify the suspect based on their footprints. However, due to the complexity of crime scenes, shoeprint images often suffer from problems such as blending of the shoe sole patterns with the background, image fragmentation, and acquisition differences, making accurate matching challenging. To address these issues, this paper proposes a shoeprint retrieval network based on a composite attention mechanism (CASN) that features a decoupled and disentangled enhanced position attention module (DEPAM) and a grouped channel attention module (GCAM). Specifically, the DEPAM module decouples highly coupled input information into salient texture information and whitened area similarity information, while the GCAM module captures long-range contextual information in the channel dimension. We introduce a spectral difference orthogonality regularization term to reduce the correlation between features, and then describe the network training method and parameter settings. Finally, by conducting comparative and ablation studies, we confirm the effectiveness of our proposed algorithm and its components, with our CASN achieving outstanding performance on the FID-300 dataset, securing top cumulative matching scores across the board except for Rank1%.

Keywords: Shoeprint image retrieval; Attention mechanism; Scene of crime

1 Introduction

Shoeprint retrieval is one of the preliminary tasks in crime scene footwear identification. Forensic investigators must first identify the type and brand of shoes worn by the suspect to collect their footprints and determine the actual perpetrator. Since shoeprint retrieval only focuses on the texture information of the footprint, footprint images are also called shoeprint images in shoeprint retrieval tasks. In most cases, criminals do not leave behind the shoes they wore during the crime, so to identify suspects based on footprints, it is necessary to first retrieve the shoes worn by the criminals through their footprint textures.

Traditional shoeprint retrieval algorithms are mainly divided into three categories based on the types of extracted features: global feature-based shoeprint retrieval algorithms,

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

region feature-based shoeprint retrieval algorithms, and interest point feature-based shoeprint retrieval algorithms.

- Global feature-based shoeprint retrieval algorithms treat shoeprint images as a whole [1]. Some researchers consider shoeprint retrieval as a geometric shape recognition problem and use geometric moment invariants to describe shoeprint images [2–4]. These methods have shown good performance on high-quality shoeprints using geometric moment invariants. Fourier transform is an excellent image representation method, especially in the field of frequency domain image analysis, so it is widely used in shoeprint retrieval [5–7]. However, the Fourier transform cannot perform local analysis. To overcome this limitation, some shoeprint image retrieval methods adopt the Gabor transform [8, 9].
- Region feature-based methods usually divide shoeprint images into different regions and then extract features from each image region [10–23]. These methods describe shoeprint images from a local perspective but do not describe the shoeprint images on a global level. These methods perform well in handling partial shoeprint images and shoe soles with periodic patterns. However, these methods may ignore the global features of shoeprint images, and some methods may fail to handle changes in image translation and rotation, resulting in recognition failures. Therefore, region feature-based methods have certain limitations in shoeprint retrieval.
- Interest point feature-based methods [24–28] are also commonly used for shoeprint image retrieval. They can be applied not only to shoeprint image retrieval, but also to many other tasks such as object recognition, classification, scene classification, and object detection. This method first detects interest points in the image and then uses local features to describe these interest points for shoeprint retrieval. These methods have good robustness and can adapt well to rotation, translation, and scale changes. Overall, interest point feature-based shoeprint retrieval algorithms have a certain degree of accuracy and robustness, but their performance may be affected by the detection of interest points.

In recent years, with the rapid development of deep learning, many tasks have been tackled using deep learning [29, 30]. More researchers have started to use convolutional neural networks to extract features from shoeprint images. This feature extraction method has stronger discriminative power when facing low-quality footprint images. Kong et al. [31] used convolutional neural networks to extract deep features of shoeprint images and then used multi-channel normalized cross-correlation methods to match these deep features. This method was tested on the publicly available FID-300 dataset [16], and the matching accuracy of the top 20% in the ranked list was about 94.0%. Zhang et al. [32] fine-tuned the parameters of the VGG-16 network using an extended shoeprint image database and then used the fine-tuned network to extract features. The test results showed that the top ten matching accuracy was about 88.7%. Ma et al. [33] proposed a new multi-part weighted convolutional neural network for shoeprint image retrieval. This method vertically divides the input image into two parts, extracts sub-features separately through a shared parameter network and finally concatenates them using a triplet loss function to measure the similarity between query images and database images. The experimental results on the FID-300 dataset showed that the accuracy of the top 10% rankings was about 89.8%.

Although the above-mentioned methods achieve certain results in extracting shoeprint features using convolutional neural networks, applying convolutional operations to the

entire image may cause the loss of some local detail information, which may affect retrieval accuracy. There are still some issues:

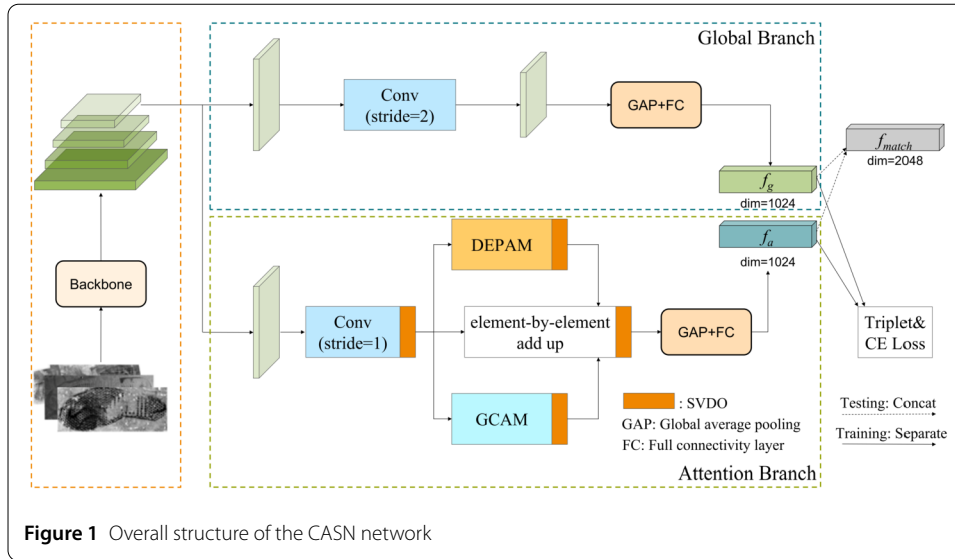
- **Insufficient local detail information:** Due to the complex background of the crime scene, shoeprint images are often contaminated with various noises, and the shoe sole patterns blend with the background. In such cases, clear local textures become crucial for retrieval. However, directly applying convolutional operations to the entire image may cause a loss of local detail information, thereby affecting retrieval accuracy.
- **Lack of long-distance correlation:** In crime scenes, there is a high probability of incomplete shoeprint images, and there is a higher similarity in local texture patterns. To retrieve the correct shoe based on incomplete scene shoeprints, it is necessary to establish long-distance dependencies between shoeprint images.
- **Feature variations:** During the collection of shoeprints at crime scenes, the acquired shoeprint images are often influenced by scale, lighting, and viewing angles. Convolutional operations result in local receptive fields, which may introduce differences in features corresponding to pixels with the same label.

Therefore, it is worth exploring how to effectively integrate the global information and local detail information of shoeprints. Based on the above issues, we propose Composite Attention Mechanism (CASN), which mainly introduces two attention modules. First, to overcome the first obstacle mentioned above, we propose Disentangled Enhanced Position Attention Module (DEPAM). Based on the decoupling idea of Disentangled Non-Local (DNL) module, the highly coupled input information in the spatial attention module is decoupled into saliency texture information and whitened region similarity information, and the fusion of the two features enhances the learning of texture and region information in the shoeprint image, reduces the computational complexity of the module, and proposes a more robust shoeprint feature. Through the DEPAM, we have achieved good performance improvement for shoeprint retrieval. Second, to address the second obstacle, we propose Grouped Channel Attention Module (GCAM), which groups the channels of features and improves the module's grasp of local information in shoeprint images through the self-attention mechanism of local channels. Then, the global information branch guides the selective fusion of grouped channel attention features to enhance the correlation of global information. This module focuses on local information while strengthening the correlation of global information, thereby further improving the network's retrieval ability. Finally, to address the third obstacle, we introduce a spectral value difference orthogonality regularization term to reduce the correlation between features, focusing more on the detailed information of shoeprints and improving the accuracy of shoeprint retrieval.

We conducted comparative experiments and a large number of ablation experiments to demonstrate the superiority of CASN in shoeprint retrieval and the importance of each module. Our approach achieved comparable results to the state-of-the-art shoeprint retrieval methods on the FID-300 dataset.

2 Methods

In this section, we introduce our CASN network. The core idea behind CASN is to leverage the advantages of attention mechanisms to enhance the focus on local features in shoeprint images while establishing long-range dependencies to improve retrieval accuracy for incomplete shoeprints.



2.1 The overall framework of the algorithm

The overall architecture of the proposed algorithm is shown in Fig. 1. The backbone of the network is ResNet-50 [34]. Before *res_conv4*, the network follows the same structure as ResNet-50. After *res_conv4*, the network branches into a global branch and an attention branch, running in parallel.

In the attention branch, the network uses the *res_conv5* layer, which is the same as in ResNet-50, but without downsampling operations. Downsampling operations are not applied here to enlarge the receptive field of convolutional layers, to preserve more detailed local features. After obtaining a feature from the *res_conv5* layer, Spectral Value Difference Orthogonality (SVDO) is applied to process it, resulting in feature z_a . Which is then fed into both the DEPAM and CGAM attention modules, with SVDO applied after each attention module. Finally, the outputs of the two attention modules are element-wise added to the input feature z_a for feature fusion. They are collectively passed through a global average pooling layer, resulting in a 1024-dimensional feature vector f_a .

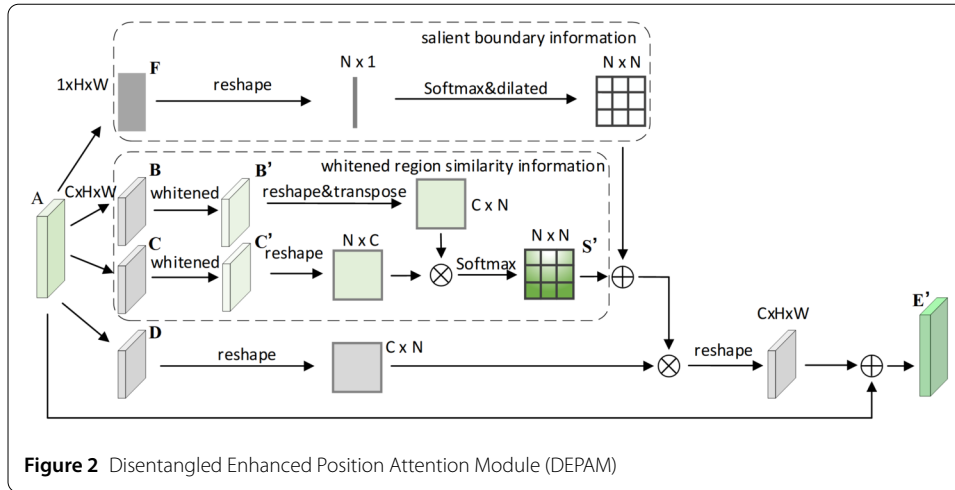
In the global branch, the complete *res_conv5* layer is retained. The feature obtained here is passed through a global average pooling layer and then a dimension reduction layer, resulting in a 1024-dimensional feature vector f_g . The global branch aims to capture global contextual information in addition to the features from the attention branch.

During training, the network converges by individually feeding $\{f_a, f_g\}$ into the loss function. During testing, to obtain stronger discriminative features, the global branch feature f_g and attention branch feature f_a are concatenated, resulting in a 2048-dimensional vector f_{match} for similarity measurement. Formally,

$$f_{match} = \text{Concat}(f_a, f_g). \quad (1)$$

2.2 Disentangled enhanced position attention module

Position attention modules have been proven to be highly effective in tasks such as semantic segmentation and person re-identification. However, directly incorporating a spatial attention module into shoeprint retrieval tasks does not fully exploit its potential performance. Firstly, crime scene backgrounds are often complex and contain various distracting



factors. After preprocessing, the shoeprint images are converted into binary images, which leads to the loss of hierarchical semantic information, including shoeprints, backgrounds, and noise. This requires the network to have stronger region learning capabilities to differentiate relevant texture information effectively. Secondly, in shoeprint retrieval tasks, texture information is primarily used for matching. Therefore, boundary information in the images is crucial. The attention module should focus more on the texture information within the shoeprint images. To address these issues, this section proposes a Disentangled Enhanced Position Attention Module (DEPAM), as illustrated in Fig. 2.

The Position Attention Module (PAM) was initially proposed by Fu et al. [35] to capture long-range contextual information to obtain discriminative features, as it is crucial for image understanding. Chen et al. [36] discovered in the field of pedestrian re-identification that using the spatial attention module can capture and aggregate semantically relevant pixels in space. This enables the network to correct misalignments of pedestrians, eliminate background disturbances, and focus on the local recognition part of the body appearance, thereby achieving better matching.

Specifically, the calculation process of the PAM in the network is as follows: Given a local feature $A \in \mathbb{R}^{C \times H \times W}$, first use a convolutional layer to generate two new feature maps B and C , where $\{B, C\} \in \mathbb{R}^{C \times H \times W}$. Then, reshape them into a dimension of $\mathbb{R}^{C \times N}$, where $N = H \times W$ is the number of pixels in the image. Next, perform matrix multiplication on the transpose of C and B , and apply a softmax layer to calculate the spatial attention matrix $S \in \mathbb{R}^{N \times N}$:

$$s_{ji} = \frac{\exp(B_i \cdot C_j)}{\sum_{i=1}^N \exp(B_i \cdot C_j)}, \quad (2)$$

where s_{ji} represents the magnitude of the influence of the position i on the position j . The similarity between the feature representations of two positions determines the level of correlation between them. Meanwhile, feature A is also input into a convolutional layer to generate a new feature map $D \in \mathbb{R}^{C \times H \times W}$, which is then transformed into dimension $\mathbb{R}^{C \times N}$. The matrix multiplication between D and the transpose of S is performed, and the result is further transformed into dimension $\mathbb{R}^{C \times H \times W}$. Finally, the obtained features are scaled by a learnable parameter and element-wise added with the original feature A to

obtain the final output feature $E \in \mathbb{R}^{C \times H \times W}$:

$$E_j = \alpha \sum_{i=1}^N (s_{ji} D_i) + A_j, \quad (3)$$

where the initial weight is set to 0 and gradually allocated more weights through learning [25]. According to Equation (3), it can be inferred that the output feature at each position is the weighted sum of the features from all positions and the original feature.

The idea of decoupling attention stems from the DNL [37] module proposed by Yin et al. Specifically, salient boundary information effectively models region boundaries, while whitened region similarity information learns the relationships between pixels within a region. Due to the strong coupling between these two types of information in the input non-local self-attention module, solely utilizing one type often results in the loss of contribution from the other. By decoupling these two pieces of information, they can be treated as separate branches and combined within the model. This approach simultaneously enhances both region boundary modeling and intra-region relationship learning. This indicates that the decoupling of information is mutually beneficial, making it necessary to adopt this decoupled information.

Based on this concept, this section proposes a decoupled enhanced spatial attention module, which decouples the information in the original spatial attention module. This not only reduces the computational complexity of the module but also allows for a stronger focus on the texture information within shoeprint images, resulting in better performance.

Specifically, for a local feature $A \in \mathbb{R}^{C \times H \times W}$, its information is decoupled as follows: for the feature maps B and C generated from A , a whitening module is applied to them, as shown in Equation (4), resulting in whitened features B' and C' . This allows for better adjustment of the original features based on the whitened region similarity, enabling the network to effectively learn intra-region relationships. Subsequently, the transposed versions of B' and C' are matrix multiplied to obtain the whitened point-to-point similarity matrix $S' \in \mathbb{R}^{N \times N}$. Formally,

$$\text{Whiten}(F) = F - \text{Avg}(F), \quad (4)$$

where $\text{Whiten}(\cdot)$ represents the operation of feature whitening, F represents the input feature, and $\text{Avg}(\cdot)$ represents global average pooling.

In addition, on top of the original structure, an extra boundary learning branch is added. It generates a feature map $F \in \mathbb{R}^{1 \times H \times W}$ through 1×1 convolution, then transforms its dimension to $\mathbb{R}^{1 \times N}$. Next, a Softmax module is applied to nonlinearly map the feature map to a distribution ranging from 0 to 1, representing salient boundary information. The dimension of the salient boundary information is expanded to $\mathbb{R}^{N \times N}$ through dimension expansion, resulting in a per-pixel salient boundary information matrix U that effectively models region boundaries. Thus, the input information in the spatial attention module is successfully decoupled into salient boundary information and whitened region similarity information. Finally, the network adds the salient boundary matrix U and the whitened region similarity matrix S' together to obtain a fused feature. Other operations within the spatial attention module are then performed, eventually yielding the output feature $E' \in \mathbb{R}^{C \times H \times W}$.

2.3 Grouped channel attention module

Building a channel attention module can model the dependencies between pairwise channels, making the channel associations with high semantic similarity stronger. However, in the context of complex crime scenes, after obtaining shoeprint images, they are generally preprocessed by converting them into binary images to enhance their texture information. However, binarization also causes a loss of some semantic information in the images, which makes it difficult for the Channel Attention Module to capture effective long-distance semantic information. As a result, its discriminability decreases significantly when dealing with images with similar footwear textures. Therefore, this section proposes GCAM that enhances the relevance of channel attention to local information while not losing the dependence on global information, thereby improving retrieval performance. Channel Attention Module (CAM) operates on the channel dimension to capture long-range contextual information. Fu et al. [35] suggest that each channel map of high-level features represents the response for a specific category, and different semantic responses are interrelated. By utilizing the correlations between these channel maps, the CAM emphasizes relevant feature maps and enhances the feature representation capability for specific semantics.

The calculation process of the CAM is as follows: Unlike the Spatial Attention Module, the channel attention matrix $X \in \mathbb{R}^{C \times C}$ directly comes from the input feature $A \in \mathbb{R}^{C \times H \times W}$. Specifically, the dimension of the feature A is first transformed to $\mathbb{R}^{C \times N}$, where $N = H \times W$ is the number of pixels in the image. Then, the matrix multiplication is performed between A and its transpose, followed by applying the softmax layer to compute the channel attention matrix $X \in \mathbb{R}^{C \times C}$. Formally,

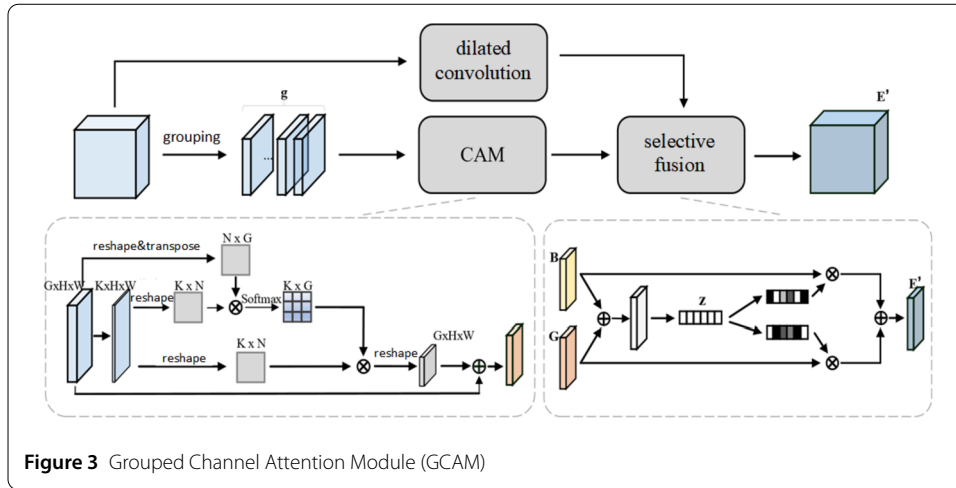
$$x_{ji} = \frac{\exp(A_i \cdot A_j)}{\sum_{i=1}^C \exp(A_i \cdot A_j)}, \quad (5)$$

where x_{ji} represents the influence of the channel i on the channel j . In addition, the transpose of X is multiplied by A , and then their resulting dimensions are transformed to $\mathbb{R}^{C \times H \times W}$. Finally, this result is multiplied with a learnable parameter β and element-wise summed with A to obtain the final output $E \in \mathbb{R}^{C \times H \times W}$. Formally,

$$E_j = \beta \sum_{i=1}^c (x_{ji} A_i) + A_j, \quad (6)$$

where β is a weight learned from 0. As shown in Equation (6), the final feature of each channel is the weighted sum of all channel features and the original features, which can model long-range semantic dependencies between feature maps. It helps improve feature distinctiveness.

As shown in Fig. 3, for an input feature $A \in \mathbb{R}^{C \times H \times W}$, it is first divided into groups by performing a grouping operation, where the features are grouped based on the number of channels into g groups, with each group having $G = C/g$ feature channels. Then, a 1×1 convolution is used to aggregate the feature channels of each group, reducing computational complexity while enhancing the learning ability of local feature channels. Each group's features become $\mathbb{R}^{K \times H \times W}$. Next, the channel attention matrix $X \in \mathbb{R}^{K \times G}$ is computed, selectively merging the aggregation centers of each group channel onto the chan-



nels of each group feature, resulting in feature groups after channel aggregation. Finally, the g groups of features are concatenated to obtain the feature $G \in \mathbb{R}^{G \times H \times W}$.

To capture the correlations between the feature channels of each group, a global information branch is introduced, which guides the fusion of the grouped channel attention features. First, to increase the effective receptive field of the feature maps and capture more contextual information, a dilated convolution with a receptive field of 5×5 is applied to feature $A \in \mathbb{R}^{C \times H \times W}$, resulting in feature B . Then, after adding feature B to G , the resulting feature undergoes global average pooling to map the feature dimension to d , generating a selection weight vector $z \in \mathbb{R}^{d \times 1}$. Afterward, a cross-channel soft attention mechanism assigns values to both the global information and the local attention information. Formally,

$$a_c = \frac{e^{A_c z}}{e^{A_c z} + e^{B_c z}}, b_c = \frac{e^{B_c z}}{e^{A_c z} + e^{B_c z}}, \quad (7)$$

$$E' = a_c \cdot B + b_c \cdot G, a_c + b_c = 1, \quad (8)$$

where $A, B \in \mathbb{R}^{C \times d}$, and a, b represent the soft attention vectors of B and G , respectively. Note that $A_c \in \mathbb{R}^{1 \times d}$ represents the row c of A , and a_c represents the element c of a . Similarly, B_c and b_c . Finally, by using different weights, the global information and local attention information are effectively fused to obtain the output feature $E' \in \mathbb{R}^{C \times H \times W}$.

2.4 Orthogonality regularization term

Orthogonality regularization is a regularization method used in deep learning. Its goal is to minimize the correlation between features in the model during training, leading to better generalization when handling different data. There are two types of orthogonality regularization: feature-wise regularization and weight-wise regularization. Feature-wise regularization can be achieved by orthogonalizing the feature vectors. For convolutional layers and fully connected layers, each feature vector can be orthogonalized to reduce its correlation. Weight-wise regularization can be done by decomposing the weight matrix into the product of two orthogonal matrices. This enforces orthogonality among the column vectors in the matrix, reducing the correlation between weights.

In this section, the orthogonal regularization method is introduced into the shoe print retrieval algorithm. By applying orthogonality regularization in the feature space, the correlation between features is reduced, allowing the model to focus more on the details of the shoe prints and improve the matching accuracy. It also helps to prevent overfitting in the network.

The specific approach used in this section is Spectral Value Difference Orthogonality (SVDO) [36]. For a feature matrix $M \in \mathbb{R}^{C \times H \times W}$, where C , H and W represent the number of channels, height, and width of features respectively, the first step is to perform a dimension transformation on M to obtain the matrix $F \in \mathbb{R}^{C \times N}$. Many orthogonal methods rely on Singular Value Decomposition (SVD) with hard orthogonality constraints to optimize F . However, SVD is computationally expensive, so soft orthogonality constraints are used instead. Many soft regularization methods [38, 39] constrain the Gram matrix of F to approximate the identity matrix under the Frobenius norm, which avoids the need for the SVD step and is differentiable. However, due to the rank deficiency of the matrix, may not satisfy this identity, resulting in biases in these regularization methods. Therefore, a regularization term based on spectral norm can effectively alleviate such biases. Additionally, orthogonality can be enhanced by directly regularizing the condition number of FF^T :

$$\beta \|k(F) - 1\|_2^2, \quad (9)$$

where β is the coefficient and $k(F)$ represents the condition number of F , defined as the ratio between the maximum and minimum singular values. Directly solving for $k(F)$ requires a complete SVD. To make it computationally tractable, Equation (9) is converted into a Spectral Value Difference Orthogonality (SVDO) regularization term:

$$\beta \|\lambda_1(FF^T) - \lambda_2(FF^T)\|_2^2, \quad (10)$$

where $\lambda_1(FF^T)$ and $\lambda_2(FF^T)$ represent the maximum and minimum eigenvalues of FF^T , respectively. In this case, autodifferentiation is used to obtain the gradient of SVDO, but this computation still involves an expensive eigenvalue decomposition (EVD). To bypass EVD, power iteration is employed to approximate the eigenvalues. Starting from a random initialization, Equation (11) is iteratively executed (typically twice):

$$p \leftarrow X_q, q \leftarrow X_p, \lambda(X) \leftarrow \frac{\|q\|}{\|p\|}, \quad (11)$$

where X refers to FF^T , where X is used to compute the maximum eigenvalue λ_1 of the Gram matrix of FF^T , and $FF^T - \lambda_1 I$ is used to compute the minimum eigenvalue λ_2 . This method can effectively calculate the SVDO regularization term.

In summary, the introduced SVDO can effectively reduce the correlation between features, avoid overfitting in the network, and improve the matching performance of the shoe print retrieval algorithm.

2.5 Loss function

The loss function used in the proposed algorithm during training consists of three components: cross-entropy loss (CE Loss), triplet loss [40], and orthogonal constraint (OF)

penalty for features. The specific expression of the loss function is shown as follows:

$$L = L_{CE} + \beta_{tr} L_{triplet} + \beta_{OF} L_{OF}, \quad (12)$$

where β_{tr} and β_{OF} are hyperparameters, and L_{OF} is the SVDO penalty term applied to the hidden features. The cross-entropy loss treats shoe sole retrieval as a classification task, aiming to minimize the gap between the model's predicted class and the true class to improve the accuracy of shoe print retrieval. The formula for cross-entropy loss is shown as follows:

$$L_{CE} = - \sum_{i=1}^N \log \frac{e^{w_{Y_i}^T f_i}}{\sum_{k=1}^C e^{w_k^T f_i}}, \quad (13)$$

where W_k is the weight vector of the class k , N is the size of the mini-batch during training, and C is the number of different shoe sole patterns in the training set.

In addition, the triplet loss aims to reduce the distance between samples of the same class and increase the distance between samples of different classes to achieve more accurate shoe sole retrieval. Its specific formula is shown as follows:

$$L_{triplet} = \sum_{i=1}^N \sum_{a,p,n} [\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \text{margin}]_+, \quad (14)$$

where N is the number of samples in the training set, a , p and n represent the anchor sample, positive sample, and negative sample, respectively. The positive sample refers to a shoe that belongs to the same pair as the anchor sample, while the negative sample refers to a shoe that is different from the anchor sample. f represents the embedding model, x_i^a , x_i^p and x_i^n represent the anchor, positive, and negative samples of the sample i , respectively. This loss function calculates the Euclidean distance between each anchor sample and its positive and negative sample and then adds a margin value. Finally, the ReLU function is applied to ensure that the loss only considers the difference between positive and negative samples. It is worth mentioning that the triplet loss used in this chapter during training is the hard-mining triplet loss. Instead of calculating the loss for all triplets, it selectively calculates the loss for the most challenging triplets. This is achieved by selecting the triplets that violate the loss function, such as choosing the negative sample with the maximum distance, to ensure that the triplet provides the maximum loss.

3 Results and discussion

We first compare CASN with other state-of-the-art algorithms on the FID-300 dataset to highlight the superiority of our proposed algorithm. We then conduct ablation experiments to analyze the various modules proposed in Sect. 2, validating their rationality and effectiveness. Next, we analyze the network's recognition ability for various image types through the visualization of retrieval results.

3.1 Training parameters and details

3.1.1 Datasets and evaluation metrics

In the training phase, we used the CS dataset [41], ITSPDB dataset [20], WVU2019 dataset [42], and some of our previously collected data as the training set. In the testing phase,

most state-of-the-art shoeprint retrieval algorithms are evaluated on the FID-300 dataset, so the subsequent experimental evaluations in this chapter are also conducted on the FID-300 dataset.

We primarily use the Cumulative Matching Score (CMS) as the evaluation metric to assess the performance of shoeprint retrieval algorithms. When comparing the cumulative matching scores of the top-ranked matches, we use the values of Rank1%, Rank5%, Rank10%, Rank15%, and Rank20% in this section.

3.1.2 Experimental implementation details

During the training process, input images are resized to a 384×144 size, and then data augmentation methods are used to expand the training set. To enhance the diversity of shoeprint samples, we employed ten data augmentation techniques, including random flipping, elastic deformation, heel and sole cropping, affine transformation, edge detection, Gaussian noise addition, random erasing, local region distortion, random pixel intensity changes, and random shifting. When an image is input to the network, the network duplicates the image 32 times and then applies these ten augmentation techniques with certain probabilities, resulting in the generation of 32 distinct shoeprint images. We employed a two-step transfer learning algorithm [43] for model fine-tuning during training. Firstly, the weights of the backbone network were frozen, and only the dimension reduction layer, classifier, and all attention modules were trained using cross-entropy loss and triplet loss for 10 epochs. Then, all layers were unfrozen, and the entire network was trained using the complete loss function for 60 epochs. The margin parameter a for the triplet loss was set to 1.2, β_{tr} was set to 0.1, and β_{OF} was set to 1×10^{-6} . The training was conducted on a single GTX 3090 GPU. The batch size was set to 64, with each batch containing 16 types of shoeprints and 4 images per type. The Adam optimizer was used, with an initial base learning rate of 3×10^{-4} , which gradually decayed to 3×10^{-5} after 30 epochs and further decayed to 3×10^{-6} after 40 epochs.

3.2 Comparison with other state-of-the-art algorithms

To further demonstrate the superiority of our CASN network in shoeprint retrieval, we compare it with 10 other shoeprint retrieval algorithms on the FID-300 dataset. These algorithms include SIFT, POC, Wu [44], Kong [45], Kong [31], Kortylewski [16], Kortylewski [17], Almaadeed [28], Wang [19], and Ma [33]. It is worth mentioning that since most of these algorithms are not open-source and the datasets they were trained on are self-constructed, the algorithm results presented here are extracted from publicly available papers. The comparison results with these algorithms are shown in Table 1.

As shown in Table 1, the proposed CASN in this chapter has achieved good performance on the FID-300 dataset. Except for Rank1%, the proposed algorithm obtains first-place results in the cumulative matching scores for the remaining metrics, and it outperforms the second-place results by more than 4%. However, the algorithm proposed in this paper has a lower cumulative matching score for Rank1% compared to the Multi-Channel Normalized Cross-Correlation method (MCNCC) proposed by Kong et al. [35]. The main reason is that the FID-300 dataset contains shoe print images with very small shoe sole patterns. MCNCC uses template matching and searches with a specific step size in both translation and rotation directions. This method is effective for small-sized images, but using normalized cross-correlation methods with multiple channels requires a considerable amount of computation.

Table 1 Performance Comparison with Other State-of-the-Art Algorithms on FID-300 Dataset

Algorithm	Cumulative Matching Score				
	Rank1%	Rank5%	Rank10%	Rank15%	Rank20%
SIFT	8.3%	18.0%	29.0%	35.6%	37.3%
POC	5.7%	14.3%	23.3%	26.3%	32.3%
Almaadeed, et al. 2015	23.7%	46.0%	59.7%	59.7%	88.0%
Kortylewski, et al. 2016	22.0%	47.5%	57.8%	67.0%	71.0%
Kortylewski. 2017	58.0%	72.0%	79.0%	81.5%	84.0%
Wang, et al. 2017	68.6%	80.9%	88.7%	93.0%	96.3%
Kong, et al. 2017	72.6%	82.3%	87.5%	91.0%	92.5%
Ma, et al. 2019	61.0%	81.4%	89.8%	-	-
Wu, et al. 2019	68.8%	81.3%	89.5%	93.0%	96.3%
Kong, et al. 2019	79.7%	86.3%	89.3%	91.3%	94.0%
CASN(Our)	76.3%	90.9%	95.3%	96.5%	98.7%

3.3 Ablation experiments

In this subsection, a series of ablation experiments will be conducted to validate the superiority of the disentangled enhanced position attention module proposed in Sect. 2.2, the group normalization for channel attention module proposed in Sect. 2.3, the orthogonal regularization method in the feature space introduced in Sect. 2.4, and the loss functions introduced in Sect. 2.5.

3.3.1 Ablation experiments on composite attention modules

In Sect. 2.2 and 2.3, we propose a disentangled enhanced position attention module and a group normalization for the channel attention module that captures long-range contextual information from both position and channel dimensions. These two modules strengthen the focus on local texture features of shoe print images and achieve better performance based on the original position attention module and channel attention module. In this subsection, five sets of ablation experiments are conducted to verify the superiority of these two modules.

Firstly, a network without the attention branches is set as the baseline. Then, two original attention modules are added to the baseline to evaluate the performance improvement brought by the attention mechanisms, denoted as B+PAM+CAM. Next, corresponding improvements are made to PAM and CAM based on B+PAM+CAM, denoted as B+DEPAM+CAM and B+PAM+GCAM, respectively. Finally, the decoupled enhanced position attention module and the channel attention module proposed in this chapter are combined to explore the performance improvement achieved by their fusion, denoted as B+DEPAM and B+GCAM. Under the same experimental settings, all five networks are trained separately and evaluated on the FID-300 dataset. The experimental results are shown in Table 2, with the best result highlighted.

From the results in Table 2, it can be observed that the addition of the PAM and CAM attention modules on top of the baseline significantly improves the performance of the network. This indicates that the attention mechanisms allow the network to focus more on the local texture features of shoe print images. Furthermore, when comparing B+DEPAM+CAM, B+PAM+GCAM, and B+PAM+CAM, it can be seen that both B+DEPAM+CAM and B+PAM+GCAM achieve comprehensive performance improvements. This suggests that both decoupling PAM and applying group normalization to CAM are beneficial for shoe print retrieval tasks. Lastly, B+DEPAM+GCAM obtains the

Table 2 Ablation experiments on composite attention modules

Method	Cumulative Matching Score				
	Rank1%	Rank5%	Rank10%	Rank15%	Rank20%
Baseline	71.6%	83.6%	90.7%	95.6%	97.6%
B+PAM+CAM	73.5%	86.5%	92.8%	96.2%	98.2%
B+DEPAM+CAM	75.2%	89.7%	94.7%	96.5%	98.7%
B+PAM+GCAM	74.9%	89.4%	94.9%	96.3%	98.6%
B+DEPAM+GCAM	76.3%	90.9%	95.3%	96.5%	98.7%

Table 3 Ablation experiments on orthogonal regularization

Method	Cumulative Matching Score				
	Rank1%	Rank5%	Rank10%	Rank15%	Rank20%
Baseline	76.3%	90.9%	95.3%	96.5%	98.7%
B-OF_p	75.1%	90.5%	94.7%	96.3%	98.7%
B-OF_c	75.6%	90.4%	94.9%	96.5%	98.5%
B-OF_all	69.2%	87.6%	93.7%	92.9%	96.3%

best result in the table, indicating that the combination of these two methods is complementary and leads to further enhancement of the network.

3.3.2 Ablation experiments on orthogonality regularization

In Sect. 2.4, we introduce an orthogonal regularization method that acts on the feature space, effectively reducing the correlation between features and enhancing the network's ability to grasp shoe print details. This subsection verifies the effectiveness of the orthogonal regularization method in the feature space through four sets of ablation experiments.

Firstly, the complete CASN proposed in this chapter is used as the baseline network, denoted as Baseline. Then, to explore the effect of OF on the attention modules, OF is removed from DEPAM and GCAM respectively based on Baseline, denoted as B-OF_p and B-OF_c. Finally, all OF in the network are removed, denoted as B-OF_all. Under the same experimental settings, these four networks are trained separately and evaluated on the FID-300 dataset. The experimental results are shown in Table 3, with the best result highlighted.

From the results in Table 3, we can see that when DEPAM and GCAM remove the Orthogonal Filter (OF), their performance decreases compared to the Baseline. This indicates that OF is effective in generating informative features for the attention mechanism. Moreover, when all OFs are removed, the network's performance drops significantly. This further demonstrates that the inclusion of orthogonality regularization helps reduce the correlation between features, enabling the model to focus more on the fine details of shoeprints and improve the accuracy of matching.

3.3.3 Analysis of ablation experiments on loss functions

In Sect. 2.5, we proposed a training approach that combines cross-entropy loss with triplet loss. The triplet loss is used to further minimize the distance between samples of the same class and maximize the distance between samples of different classes. In this section, we will experimentally verify the impact of incorporating triplet loss on the performance of the network.

Firstly, CASN is taken as the baseline network. The network trained only with cross-entropy loss is referred to as the Baseline. Then, triplet loss is added to the Baseline for joint

Table 4 Ablation Experiments on Loss Functions

Method	Cumulative Matching Score				
	Rank1%	Rank5%	Rank10%	Rank15%	Rank20%
Baseline	75.1%	88.9%	94.1%	95.6%	97.9%
B+Tri	76.3%	90.9%	95.3%	96.5%	98.7%

network training, denoted as B+Tri. Under the same experimental settings, both networks are trained separately and evaluated on the FID-300 dataset for comparison.

From the results in Table 4, it can be observed that B+Tri shows improvement in all metrics compared to the Baseline model. This indicates that the addition of the triplet loss enhances the performance of the network. The results demonstrate that the triplet loss plays an important role in increasing the discriminative power of the features, leading to improved performance.

3.4 Visual analysis of results

To further validate the effectiveness of the CASN algorithm in footwear impression retrieval tasks, this section compares the CASN algorithm proposed in this paper with the MCNCC algorithm, which showed good retrieval results in Table 1, on the FID-300 dataset. The retrieval results of both algorithms on the FID-300 dataset are shown in Fig. 4. The visual retrieval results of MCNCC are sourced from its published paper [35]. Figure 4 displays the retrieval results of 6 sets of real-life shoe impressions. Each set is divided by a blue line, with the query impression on the left and the retrieval results on the right. Each algorithm provides 8 reference shoe impressions that it considers to be most similar to the query image. If there are correctly matched reference shoe impressions in the retrieval results, they will be highlighted in green.

From the figure, it can be seen that in these 6 sets of query impressions, CASN outperforms MCNCC in 4 sets by having the correct reference shoe impressions as the top result. For “img010”, the correct reference shoe impression is ranked third in the CASN retrieval results, but the front paw textures of the first three reference shoe impressions are very similar, making it difficult to distinguish. Moreover, among the 8 retrieval images provided by the CASN algorithm, 5 have very similar front paw textures, indicating that the CASN algorithm is highly sensitive to textures. For “img003”, both MCNCC and CASN failed to retrieve the correct shoe impression image in the top 8 retrieval results, indicating that the ability of the algorithm to extract effective information needs to be further improved when facing incomplete images with small differences between the shoe impression texture and background.

4 Conclusion and future work

We propose a framework CASN for shoeprint retrieval tasks in criminal scenarios. Firstly, the DEPAM is proposed to decouple the information in the original position attention module and improve the effect of shoeprint image texture learning and intra-region relationship learning. Secondly, the GCAM is proposed to group feature channels and use the self-attention mechanism of local channels to enhance the module’s grasp of local information in shoeprint images, and selectively fuse under the guidance of the global information branch to strengthen the relevance of global information. Then, an SVDO regularization term is introduced to reduce the correlation between features, making the

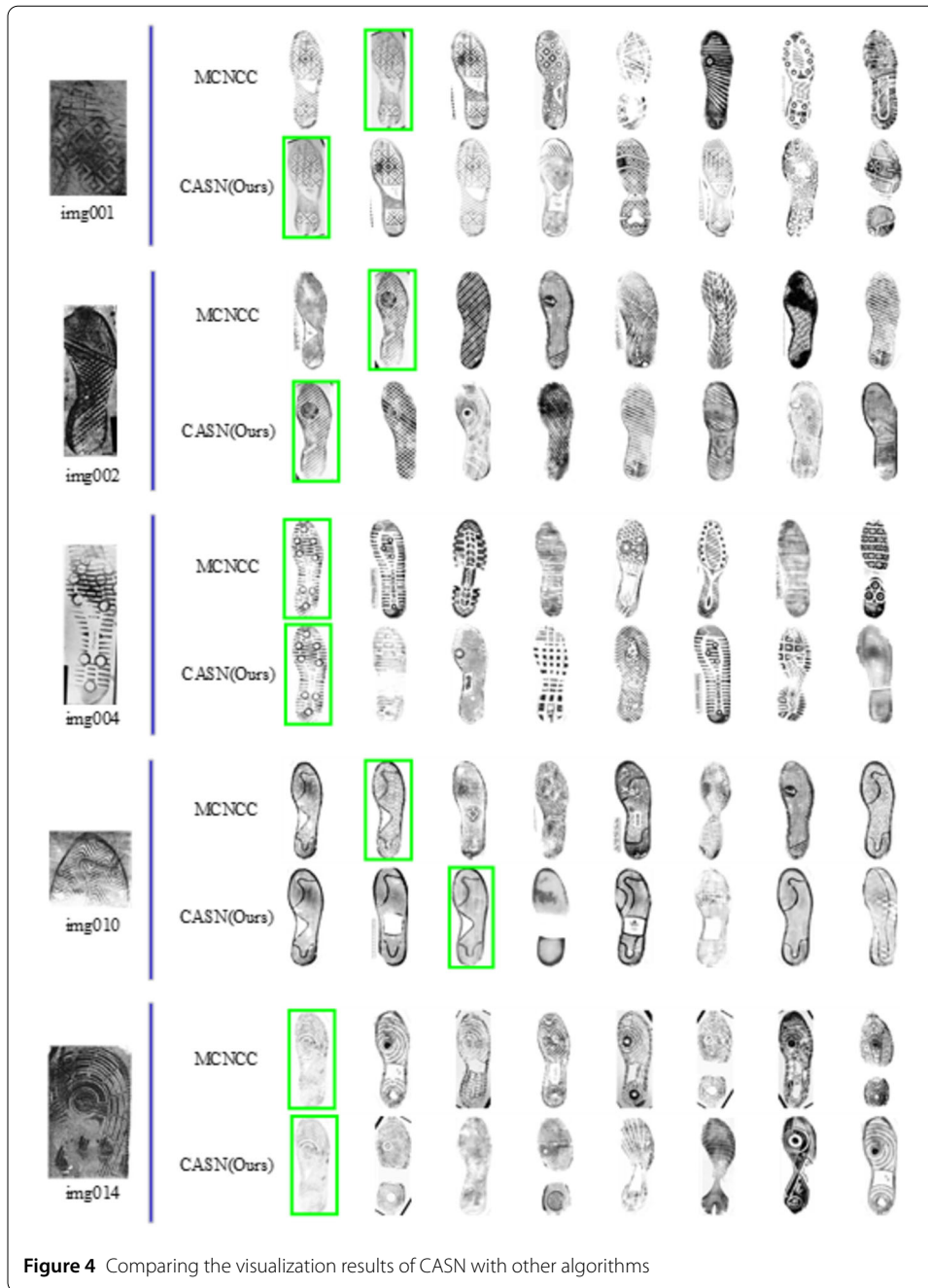


Figure 4 Comparing the visualization results of CASN with other algorithms

model more focused on the details of shoeprints and improving the accuracy of shoeprint retrieval. We conducted multiple ablation experiments on various designed modules and introduced methods, and the final experimental results demonstrate that these modules and methods are reasonable and effective. Experimental results on the FID-300 dataset demonstrated the good performance of our methods.

In future work, research on shoeprint retrieval algorithms should focus on local feature-based shoeprint retrieval algorithms. Currently, most footprint images in crime scenes are incomplete and degraded. The shoeprint retrieval algorithms that use CNN to extract global features usually produce unsatisfactory results when facing such images. Therefore,

it is possible to detect and locate incomplete images first, determine their approximate locations, and then use local features of reference images for matching retrieval to improve the accuracy of retrieval.

Acknowledgements

We would like to thank the reviewers for their time spending and the valuable comments.

Author contributions

The main idea of this paper was proposed by Liman Liu, and Liman Liu conducted the methodological design. Yifei Han prepared the manuscript initially and implemented the simulations using the software. Jinshan Tian conducted a visual study and analysis. Han Wu completed method validation. En Yu, Qiling Tang, and Wenbing Tao supervised and edited the writing. All authors read and approved the final manuscript.

Funding information

This work was supported by the National Natural Science Foundation of China [grant number 61976227, 62176096 and 62076257].

Data availability

The datasets used during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

¹Key Laboratory of Cognitive Science, State Ethnic Affairs Commission, Hubei Provincial Key Laboratory of Medical Information Analysis and Tumor Diagnosis and Treatment, School of Biomedical Engineering, South-Central Minzu University, Wuhan, Hubei, 430074, China. ²National Key Laboratory of Science and Technology on Multi-spectral Information Processing, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, Hubei, 430074, China.

Received: 30 March 2024 Accepted: 28 October 2025 Published online: 20 November 2025

References

1. Bouridane, A., Alexander, A., Nibouche, M., Crookes, D.: Application of fractals to the detection and classification of shoeprints. In: Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101), vol. 1, pp. 474–4771 (2000). <https://doi.org/10.1109/ICIP.2000.900998>
2. AlGarni, G., Hamiane, M.: A novel technique for automatic shoeprint image retrieval. *Forensic Sci. Int.* **181**(1–3), 10–14 (2008). <https://doi.org/10.1016/j.forsciint.2008.07.004>
3. Khotanzad, A., Hong, Y.H.: Invariant image recognition by Zernike moments. *IEEE Trans. Pattern Anal. Mach. Intell.* **12**(5), 489–497 (1990). <https://doi.org/10.1109/34.55109>
4. Wei, C.-H., Gwo, C.-Y.: Alignment of core point for shoeprint analysis and retrieval. In: 2014 International Conference on Information Science, Electronics and Electrical Engineering, vol. 2, pp. 1069–1072 (2014). <https://doi.org/10.1109/InfoSEE.2014.6947833>
5. Huynh, C., Chazal, P., McErlan, D., Reilly, R.B., Hannigan, T.J., Fleury, L.M.: Automatic classification of shoeprints for use in forensic science based on the Fourier transform. In: Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429), vol. 3, p. 569 (2003). <https://doi.org/10.1109/ICIP.2003.1247308>
6. Cervelli, F., Dardi, F., Carrato, S.: A texture recognition system of real shoe marks taken from crime scenes. In: 2009 16th IEEE International Conference on Image Processing (ICIP), pp. 2905–2908 (2009). <https://doi.org/10.1109/ICIP.2009.5413417>
7. Cervelli, F., Dardi, F., Carrato, S.: A translational and rotational invariant descriptor for automatic footwear retrieval of real cases shoe marks. In: 2010 18th European Signal Processing Conference, pp. 1665–1669 (2010)
8. Patil, P.M., Kulkarni, J.V.: Rotation and intensity invariant shoeprint matching using Gabor transform with application to forensic science. *Pattern Recognit.* **42**(7), 1308–1317 (2009). <https://doi.org/10.1016/j.patcog.2008.11.008>
9. Li, X., Wu, M., Shi, Z.: The retrieval of shoeprint images based on the integral histogram of the Gabor transform domain. In: Shi, Z., Wu, Z., Leake, D., Sattler, U. (eds.) *Intelligent Information Processing*, vol. VII, pp. 249–258. Springer, Berlin (2014)
10. Tang, Y., Srihari, S.N., Kasiviswanathan, H., Corso, J.J.: Footwear print retrieval system for real crime scene marks. In: Sako, H., Franke, K.Y., Saitoh, S. (eds.) *Computational Forensics*, pp. 88–100. Springer, Berlin (2011)
11. Tang, Y., Srihari, S.N., Kasiviswanathan, H.: Similarity and clustering of footwear prints. In: 2010 IEEE International Conference on Granular Computing, pp. 459–464 (2010). <https://doi.org/10.1109/GrC.2010.175>
12. Tang, Y., Kasiviswanathan, H., Srihari, S.N.: An efficient clustering-based retrieval framework for real crime scene footwear marks. *Int. J. Granul. Comput. Rough Sets Intell. Syst.* **2**(4), 327–360 (2012)
13. Pavlou, M., Allinson, N.M.: Automatic extraction and classification of footwear patterns. In: Corchado, E., Yin, H., Botti, V., Fyfe, C. (eds.) *Intelligent Data Engineering and Automated Learning – IDEAL 2006*, pp. 721–728. Springer, Berlin (2006)
14. Pavlou, M., Allinson, N.M.: Automated encoding of footwear patterns for fast indexing. *Image Vis. Comput.* **27**(4), 402–409 (2009). <https://doi.org/10.1016/j.imavis.2008.06.003>

15. Kortylewski, A., Albrecht, T., Vetter, T.: Unsupervised footwear impression analysis and retrieval from crime scene data. In: Jawahar, C.V., Shan, S. (eds.) *Computer Vision - ACCV 2014 Workshops*, pp. 644–658. Springer, Cham (2015)
16. Kortylewski, A., Vetter, T.: Probabilistic compositional active basis models for robust pattern recognition. *British Machine Vision Conference* (2016). <https://api.semanticscholar.org/CorpusID:1654127>
17. Kortylewski, A.: Model-based image analysis for forensic shoe print recognition (2017). <https://api.semanticscholar.org/CorpusID:116161335>
18. Wang, X., Sun, H., Yu, Q., Zhang, C.: Automatic shoeprint retrieval algorithm for real crime scenes. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) *Computer Vision - ACCV 2014*, pp. 399–413. Springer, Cham (2015)
19. Wang, X., Zhang, C., Wu, Y., Shu, Y.: A manifold ranking based method using hybrid features for crime scene shoeprint retrieval. *Multimed. Tools Appl.* **76**, 21629–21649 (2017)
20. Alizadeh, S., Kose, C.: Automatic retrieval of shoeprint images using blocked sparse representation. *Forensic Sci. Int.* **277**, 103–114 (2017)
21. Tang, C., Dai, X.: Automatic shoe sole pattern retrieval system based on image content of shoeprint. In: 2010 International Conference on Computer Design and Applications 4, p. 4–6024605 (2010)
22. Ghouti, L., Bouridane, A., Crookes, D.: Classification of shoeprint images using directional filterbanks. In: 2006 IET International Conference on Visual Information Engineering, pp. 167–173 (2006)
23. Alizadeh, S., Jond, H.B., Nabyev, V.V., Kose, C.: Automatic retrieval of shoeprints using modified multi-block local binary pattern. *Symmetry* **13**(2) (2021). <https://doi.org/10.3390/sym13020296>
24. Li, Z., Wei, C.-H., Li, Y., Sun, T.: Research of shoeprint image stream retrieval algorithm with scale-invariance feature transform. In: 2011 International Conference on Multimedia Technology, pp. 5488–5491 (2011)
25. Wang, H., Fan, J., Li, Y.: Research of shoeprint image matching based on sift algorithm. *J. Comput. Methods Sci. Eng.* **16**(2), 349–359 (2016)
26. Nibouche, O., Bouridane, A., Crookes, D., Gueham, M., Laadjel, M.: Rotation invariant matching of partial shoeprints. In: 2009 13th International Machine Vision and Image Processing Conference, pp. 94–98 (2009)
27. Su, H., Crookes, D., Bouridane, A., Gueham, M.: Local image features for shoeprint image retrieval. In: *Proceedings of the British Machine Vision Conference*, pp. 1–10 (2007). <https://doi.org/10.5244/C.21.38>
28. Al-Maadeed, S.A., Bouridane, A., Crookes, D., Nibouche, O.: Partial shoeprint retrieval using multiple point-of-interest detectors and sift descriptors. *Integr. Comput.-Aided Eng.* **22**(1), 41–58 (2015)
29. Khlongkhai, P., Chayantrakom, K., Kanbua, W.: Application of a deep learning technique to the problem of oil spreading in the Gulf of Thailand. *Adv. Differ. Equ.* **2019**(1), 306 (2019). <https://doi.org/10.1186/s13662-019-2241-y>
30. Kröpfel, F., Maier, R., Peterseim, D.: Operator compression with deep neural networks. *Adv. Cont. Discr. Models* **2022**(1), 29 (2022). <https://doi.org/10.1186/s13662-022-03702-y>
31. Kong, B., Supančič, J., Ramanan, D., Fowlkes, C.C.: Cross-domain image matching with deep feature maps. *Int. J. Comput. Vis.* **127**(11), 1738–1750 (2019). <https://doi.org/10.1007/s11263-018-01143-3>
32. Zhang, Y., Fu, H., Dellandrea, E., Chen, L.: Adapting convolutional neural networks on the shoeprint retrieval for forensic use. In: Zhou, J., Wang, Y., Sun, Z., Xu, Y., Shen, L., Feng, J., Shan, S., Qiao, Y., Guo, Z., Yu, S. (eds.) *Biometric Recognition*, pp. 520–527. Springer, Cham (2017)
33. Ma, Z., Ding, Y., Wen, S., Xie, J., Jin, Y., Si, Z., Wang, H.: Shoe-print image retrieval with multi-part weighted cnn. *IEEE Access* **7**, 59728–59736 (2019)
34. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). <https://doi.org/10.1109/CVPR.2016.90>
35. Fu, J., Liu, J., Tian, H., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3141–3149 (2019)
36. Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z.: Abd-net: attentive but diverse person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8350–8360 (2019)
37. Yin, M., Yao, Z., Cao, Y., Li, X., Zhang, Z., Lin, S., Hu, H.: Disentangled non-local neural networks. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) *Computer Vision - ECCV 2020*, pp. 191–207. Springer, Cham (2020)
38. Xie, D., Xiong, J., Pu, S.: All you need is beyond a good init: exploring better solution for training extremely deep convolutional neural networks with orthonormality and modulation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5075–5084 (2017)
39. Xu, H., Wang, Z., Yang, H., Liu, D., Liu, J.: Learning simple thresholded features with sparse support recovery. *IEEE Trans. Circuits Syst. Video Technol.* **30**(4), 970–982 (2020). <https://doi.org/10.1109/TCSVT.2019.2901713>
40. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823 (2015)
41. Richetelli, N., Lee, M.C., Lasky, C., Gump, M.E., Speir, J.: Classification of footwear outsole patterns using Fourier transform and local interest points. *Forensic Sci. Int.* **275**, 102–109 (2017)
42. Lin, E.-T., DeBat, T., Speir, J.: A simulated crime scene footwear impression database for teaching and research purposes. *J. Forensic Sci.* **67**(2), 726–734 (2022)
43. Costilla-Reyes, O., Vera-Rodriguez, R., Scully, P., Ozanyan, K.B.: Analysis of spatio-temporal representations for robust footprint recognition with deep residual neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(2), 285–296 (2019). <https://doi.org/10.1109/TPAMI.2018.2799847>
44. Wu, Y., Wang, X., Nankabirwa, N.L., Tao, Z.: Losgrs: learned opinion score guided shoeprint retrieval. *IEEE Access* **7**, 55073–55089 (2019)
45. Kong, B., Supančič, J., Ramana, D., Fowlkes, C.: Cross-domain forensic shoeprint matching. In: *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 1–12 (2017). <https://doi.org/10.5244/C.31.188>

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.