

A Survey of Crime Scene Investigation Image Retrieval Using Deep Learning

Ying Liu, Aodong Zhou, Jize Xue[✉], Zhijie Xu

Abstract: Crime scene investigation (CSI) image is key evidence carrier during criminal investigation, in which CSI image retrieval can assist the public police to obtain criminal clues. Moreover, with the rapid development of deep learning, data-driven paradigm has become the mainstream method of CSI image feature extraction and representation, and in this process, datasets provide effective support for CSI retrieval performance. However, there is a lack of systematic research on CSI image retrieval methods and datasets. Therefore, we present an overview of the existing works about one-class and multi-class CSI image retrieval based on deep learning. According to the research, based on their technical functionalities and implementation methods, CSI image retrieval is roughly classified into five categories: feature representation, metric learning, generative adversarial networks, autoencoder networks and attention networks. Furthermore, We analyzed the remaining challenges and discussed future work directions in this field.

Keywords: crime scene investigation (CSI) image; image retrieval; deep learning

1 Introduction

Crime scene investigation (CSI) image is generally generated by the public police during criminal investigation. And with the advent of the new detection technologies and the need for public security, the image of the criminal investigation has been explosively growing [1]. To extract valuable information from CSI image, early methods based on artificial operator are used to recognize and detect the images involved in various cases, however, such methods are time-consuming and are tend to make mistakes because of visual fatigue, subjective intention and percep-

tual bias. To avoid the problem, the scheme based on CSI image retrieval is proposed, which uses the similarity measure to the query image to acquire a ranking list of CSI images in a database. In this way, a CSI image can not only be used to facilitate the police to quickly obtain key clues, identify suspects or items but also as an exhibit for bringing criminal charges, and accelerate the process of crime detection [2, 3].

The existing CSI image retrieval can be classified into two categories: CSI image retrieval methods based on traditional features and CSI retrieval image methods based on deep learning features. Traditional features include color, texture, shape and spatial location low-level visual features. which are also termed as the hand-engineered feature, and it's worth noting that the traditional methods require selecting the appropriate feature combinations for image processing and interpretation based on task requirements. In early studies, researchers mostly utilized manually designed features for one-class and multi-class CSI image retrieval [2, 4-23].

For different image retrieval tasks, the

Manuscript received Dec. 13, 2023; revised Jan. 20, 2024; accepted Feb. 15, 2024. The associate editor coordinating the review of this manuscript was Dr. Xudong Zhao. This work was supported by the National Natural Science Foundation of China (No. 62301423) and Special Scientific Research Plan Project of Shaanxi Provincial Department of Education (No. 23JK0671).

Ying Liu, Aodong Zhou and Jize Xue are with Xi'an University of Posts and Telecommunications (XUPT), Xi'an 710121, China.

Zhijie Xu is with University of Huddersfield, Huddersfield HD1 3DH, UK.

✉ Corresponding author. Email: Jize.Xue@xupt.edu.cn

DOI: [10.15918/j.jbit1004-0579.2023.152](https://doi.org/10.15918/j.jbit1004-0579.2023.152)

researchers have used the various feature description. For example, for shoeprint image retrieval, the feature extraction techniques such as scale-invariant feature transform (SIFT) [4, 5], Gabor transform [6], Fourier transform [7], Wavelet Fourier-Mellin transform [8–10], local binary pattern (LBP) [11] are usually employed. For tattoo image retrieval, researchers not only have used the SIFT methods [12, 13], but also have designed the color histogram and correlation graph [14–16]. Considering that the color and edge knowledges of license plate, the researchers utilized techniques based on edge detection [17, 18] and color feature analysis [19] to match and identify license plate images. By employing the advantage of the Gabor filtering for depicting of image texture information, the researchers have employed the filtering [20] to extract features such as the direction and position of hair. The researchers conducted tire pattern image analysis by employing Radon transform and energy distribution-based curvelet transform [21], histogram of oriented gradients (HOG) and dominant gradient (DG) [2], as well as adaptive weighted feature fusion based on discrete wavelet transform (DWT) and LBP [22]. And, Liu et al. [23] have proposed a multi-class CSI image retrieval algorithm that combines texture features in the discrete cosine transform (DCT) domain, generalized search tree (GIST) descriptors, and fused features of HSV color histograms.

Last decade, inspired by the deep learning, it is observed that the feature representation has shifted from hand-engineering to learning-based schemes. Although image retrieval techniques based on deep learning have become mature, their application in CSI image datasets is still in the progressive development stage. The mathematical definition of the deep learning-based CSI image retrieval technique in question can be formulated as follows: Given a query image Q and a CSI image database $D = \{d_1, d_2, \dots, d_n\}$, the objective is to utilize a deep learning model to construct an image retrieval system that can find

images in the database D that are similar to the query image Q . $f(I)$ is the function of the deep learning model that converts an image I into a feature vector, where $f(I)$ is a high-dimensional vector serving as the representation of image I . The goal of CSI image retrieval is to compute the similarity between the query image Q and each image in the database D , as $\text{Similarity}(Q, d_i) = \text{Similarity}(f(Q), f(d_i))$, then sort by similarities, and return retrieved relevant images.

How to utilize deep learning techniques to quickly search for targets in large-scale CSI image datasets has become a hot research topic among scholars. Thus, aiming to facilitate scholarly research, this paper summarizes the one-class and multi-class CSI image datasets currently used by researchers. It provides an overview of recent developments in deep learning-based CSI image retrieval techniques, based on their technical functionalities and implementation methods, which can be roughly categorized into five classes: feature representation, metric learning, generative adversarial networks, autoencoder networks, and attention networks. Additionally, considering that the practical needs of the public security industry, we analyzes the development trends of CSI image retrieval technology and discusses future research directions in this field.

2 Datasets and Evaluation Metrics

2.1 One-Class CSI Image Datasets

In recent years, the scholars have been conducting indepth research in this field, at the same time which puts forward higher demands for datasets. According to literature review, various types of CSI image datasets have been established and utilized for academic research, in which most of the datasets derived from actual data in specific industries, and partial datasets are self-constructed data. The main datasets include license plate image datasets, fingerprint image datasets, shoeprint image datasets, tattoo

image datasets [24–47], etc. Please refer to Tab. 1, Tab. 2, Tab. 3 and Tab. 4 for more details, and sample images are illustrated in Fig. 1.

2.2 CIIP-CSI Image Dataset

The Center for Image and Information Processing (CIIP) at Xi’an University of Posts and Telecommunications, in collaboration with the Ministry of Public Security, has established a research platform, which has been dedicated to research in the field of CSIR over the past years. And, CIIP has obtained a substantial amount of actual crime scene investigation images from the

public security system. This dataset, is known as the CIIP Crime Scene Investigation Image Dataset (CIIP-CSID), which comprises 19 363 images from real cases and is categorized into 17 primary classes and 52 secondary subclasses. Tab. 5 presents the primary classes and the corresponding number of images included in each category within CIIP-CSID, and Fig. 2 illustrates partial examples from CIIP-CSID.

2.3 Evaluation Metrics

In order to judge the performance of CSI image retrieval approaches, precision, recall and F-score

Tab. 1 License plate image datasets

Institution	Dataset	Size	Public
The University of Science and Technology of China	CCPD[28]	250k	Yes
The University of Science and Technology of China	ChineseLP[37]	411	Yes
National Taiwan University of Science and Technology	AOLP[33]	2 049	Yes
Caltech	Caltech[32]	126	Yes
University of Technology Brno, Czech Republic	ReId[35]	76k	Yes
Federal University of Paraná, Curitiba, Brazil	RodoSol-ALPR[36]	20k	Yes
Federal University of Paraná, Curitiba, Brazil	UFPR-ALPR[29]	4 500	Yes
Smart Surveillance Interest Group, Belo Horizonte, Brazil	SSIG[31]	2 000	Yes
University of Zagreb Faculty of Electrical Engineering and Computing	Zemris[34]	510	Yes

Tab. 2 Fingerprint image datasets

Institution	Dataset	Size	Public
Changchun University of Science and Technology,China	[30]	40k	No
Coventry University, Faculty of Engineering	SOCOFing[40]	6 000	Yes
NIST Data Discovery	NIST SD[42,43]	\	Yes
Peking University,China	FVC2000 DB2a/3a[44]	800/800	No
Michigan State University, American	MSP[45]	15 597 subjects	No (applicable)

Tab. 3 Shoeprint image datasets

Institution	Dataset	Size	Public
University of Basel Basel, Switzerland	FID-300[38]	1 175+300	Yes
West Virginia University	CS[5]	100	Yes
Capital Normal University,China	[47]	2 000	No
Xi’an University of Posts and Telecommunications,China	SPID[46]	10k	No

Tab. 4 Tattoo image datasets

Institution	Dataset	Size	Public
Image Group,NIST-Information Access Division National	Tatt-C[39]	16 716	No
Nanyang Technological University, Singapore	Flickr[25]	10k	No
University of Zagreb Faculty of Electrical Engineering and Computing	DeMSI[27]	890	No
Advanced Technologies Application Center	BIVTatt[26]	210	Yes
Visual and Information Processing Lab, Chinese Academy of Sciences	WebTattoo[24]	30w	Yes

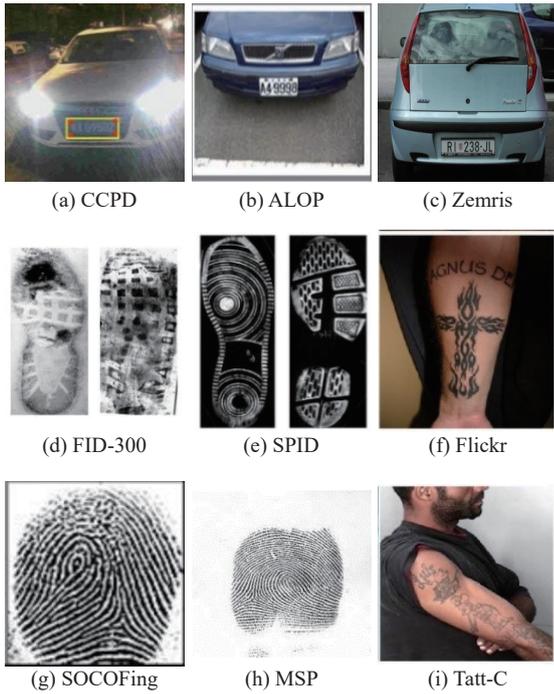


Fig. 1 Samples of CSI image datasets

Tab. 5 CIIP-CSI image dataset

Category	Number	Category	Number
Biological evidence	614	Plan	699
Bloodstain	597	Shoeprint	1717
Cars	599	Skin	1133
Distant view	1200	Tattoo	2828
Photograph		Tools for crime	971
Door	600	Purpose	5597
Fingerprint	741	Tyre pattern	361
Indoor scene	608	Tyre indentation	610
Physical evidence	454	Window	
Other	34		

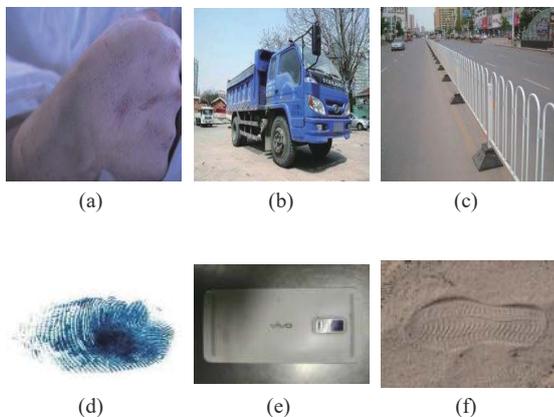


Fig. 2 Samples of CIIP-CSI image dataset: (a) biological evidence; (b) cars-truck; (c) distant view photograph; (d) fingerprint; (e) physical evidence; (f) shoeprint

are the common evaluation metrics. The precision is defined as the percentage of correctly retrieved images out of the total number of retrieved images. The recall measures the performance of an image retrieval system by calculating the percentage of correctly retrieved images out of the total relevant images present in the dataset. F-score is calculated based on the harmonic mean of precision and recall. In addition, The mean average precision (MAP) is also used for evaluating CSI image retrieval methods, which is calculated by computing the precision (P) at each position of the retrieved images for each query image, then calculating the average precision (AP) for each query image over all positions of the retrieved images, and finally computing the mean average precision (MAP) over all query images. The formula for calculating MAP is as follows:

$$MAP = \frac{\sum_{i=1}^Q AP@N(i)}{Q} \quad (1)$$

where Q is the number of query images and N is the total number of retrieved images. The precision (P) is calculated using the following formula:

$$P = \frac{S}{K} \quad (2)$$

where K is the top K retrieved images, S is the number of relevant images correctly retrieved out of the top K retrieved images, and $P@K$ represents the probability of correctly retrieving relevant images out of the top K retrieved images. $AP@K$ represents the average probability of correctly retrieving relevant images among the top K retrieved images for all query images. Some results in the literature are reported in the format of $AP@K$ and MAP.

3 Deep Learning Based CSI Image Retrieval

Traditional content-based image retrieval (CBIR) with special requirements for knowledge and expertise needs selecting suitable features

and parameters. And, it is time-consuming to compute the feature vector for each specific image. The CBIR technique based on deep learning adopts models such as convolutional neural networks (CNNs) to improve their capacity for autonomous learning and feature representation. Deep learning models can exhibiting exceptional performance in similarity matching tasks by acquiring more precise representations for underlying image features through extensive training with labeled data.

CSI image retrieval integrates CBIR with the features of the current image database, resulting in improved retrieval performance. CSI image retrieval built on deep learning can be roughly categorized based-distinct characteristics and based-technical approaches. In this paper, we categorize deep learning-based CSI image retrieval technologies into five aspects: feature representation based retrieval, metric learning based retrieval, generative adversarial networks (GAN) based retrieval, autoencoder networks based retrieval and attention networks based retrieval, as depicted in Fig. 3. The details of the

recent CSI image retrieval methods are summarized in Tab. 6.

3.1 Feature Representation Based Retrieval

The core idea of feature-based CSI image retrieval methods is to utilize deep neural networks as feature extractors. By mapping images into a feature space, the semantic similarity between images can be enhanced, which improves the similarity between the images in the feature space. Along the idea, such methods usually enables similarity matching and retrieval.

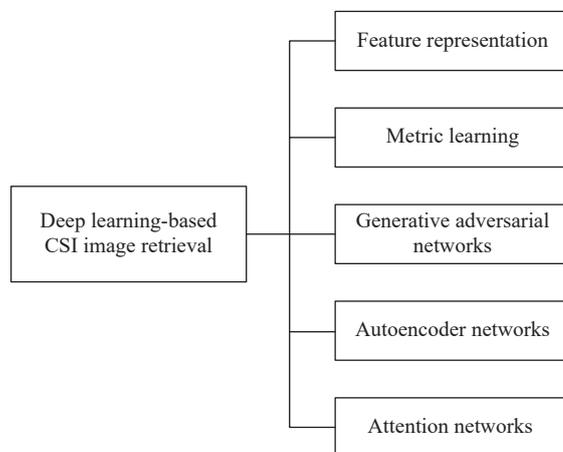


Fig. 3 Existing methods classified in 5 categories

Tab. 6 Summary of the published CSI image retrieval methods

Method	Model	Feature	Matching method	Performance	Dataset
Nicolás-Díaz et al.[26]	MobileNetV2	Conv-Layer feature	Euclidean distance	89.935%@10% 76.806%@20%	BIVTatt PinTatt
Li et al.[48]	CNN+RPN+CTC	Conv-Layer feature	\	83.80%	ALOP (RP)
Xu et al.[28]	RPnet (CNN high-low layer +ROI)	Conv-Layer feature	\	95.5%	CCPD
Di et al.[49]	CNN+Siamese	Conv-Layer feature	Triplet loss	56.9%@10	Tatt-C
Wen et al.[50]	CNN+NCC	Conv-Layer feature	Normalized cross-correlation	82%	FID-300
Liu et al.[51]	Multi CNN +PCA	Fusion feature	5597	92.33%	CIIP-CSID
Zhang et al.[52]	CNN Multi-layer	Fusion feature	SVM	93.67%	CIIP-CSID
Ma et al.[46]	Multi-part weighted CNN	Fusion feature	Triplet loss	75.11%@10% 89.83%@10%	SPID FID-300
Kong et al.[53]	CNN	Conv-Layer feature	Multi-channel normalized cross-correlation	86.33%@5%	FID-300
Jiawang et al.[54]	GAN+Triplet	Deep features	Triplet loss	77%–82%	Authors-created
Cao et al.[44]	Convolutional autoencoder	Conv-Layer feature	Euclidean distance	65.7%@1	NIST SD27
Nicolás-Díaz et al.[55]	CNN+WAP	Conv-Layer feature	Euclidean distance	93.571%@10% 81.244%@10%	BIVTatt PinTatt
Sun et al.[30]	APFI (CNN+SE)	Deep features	Angular distance	98.9%@20%	NIST SD4
Li et al.[56]	DKD (SA+CNN+Hash)	Deep features	Hamming distance	80.49%@1% 75.63%@1%	FID-300 SPID
Liu et al.[57]	CNN+STN	Deep features	Euclidean distance	0.819 (MAP)	Authors-created

In the early studies, the researchers have employed pre-trained CNNs as feature extractors, including popular architectures such as VGGNet [58], ResNet [59], Inception [60–62], AlexNet [63], etc. In general, these networks first are initially trained on large-scale image classification tasks, enabling them to learn highly discriminative representations of image features. Then, one feeds the input image into a pre-trained model, and uses the advantage of model to extract the features. These extracted features, often referred to as feature vectors or feature descriptors, capture the underlying characteristics of the image. Actually, the image data can be utilized to construct a feature vector database. In the process of the

image retrieval, the queried image is passed through the feature extraction network to obtain its corresponding feature vector. Subsequently, by evaluating the similarity between the feature vector of the queried image and those of all the images in the database, the search task is accomplished until generating a sorted list.

The similarity measurement methods commonly employed include Euclidean distance, cosine similarity, etc. The process is visually depicted in Fig. 4, illustrating the sequential steps involved in image retrieval, including feature extraction, similarity calculation, and sorting of images based on their similarity scores.

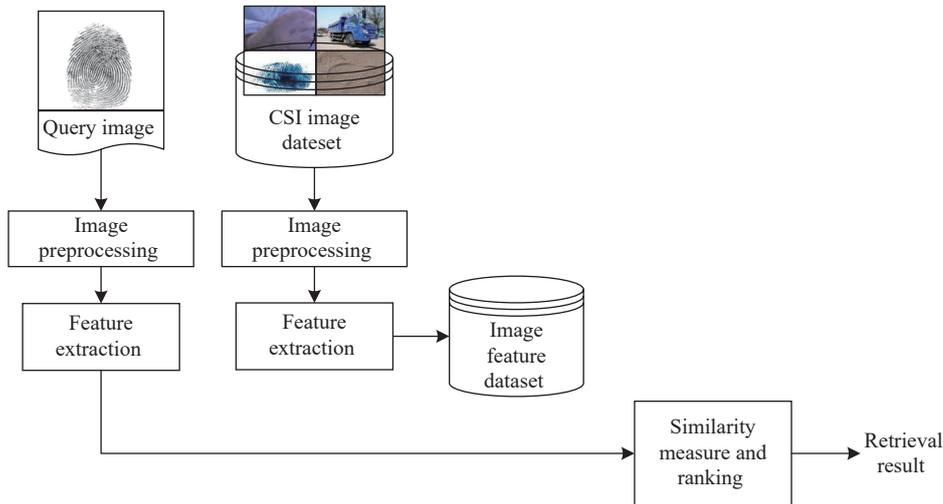


Fig. 4 The framework of CSI image retrieval based on feature representation

The feature representation based CSI image retrieval method possess the following advantages: 1) Learning semantically rich image representations enables improved similarity measurement. 2) Compact feature vectors reduce storage and computational costs. However, there are also some limitations to this approach: 1) Some information may be lost as typically only features from a subset of layers in the network are retained. 2) Limited by the generalization capability of the pre-trained network, it may perform poorly in certain specific domains or tasks. The main distinction among existing feature-based CSI image retrieval methods lies in the types of features employed. Based on the property of

these features, these methods can be broadly categorized into three major classes, as in Fig. 5, depicting the framework for feature extraction.

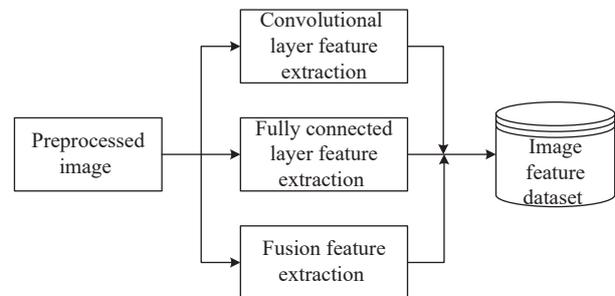


Fig. 5 The framework of feature extraction

3.1.1 Convolutional Layer Feature Extraction

In deep learning-based CSI image retrieval, pre-

trained CNNs are commonly used to extract convolutional layer features from images. Transfer learning and other approaches are employed to apply the knowledge learned from the source domain to tasks in the target domain.

In [26], the middle layer features of a neural network that trained on the large public dataset ImageNet were utilized as the descriptor for the tattoo image, without any fine-tuning. In [48], the lower-level CNN features of license plate images were extracted by enhancing the VGG16 model as a feature extractor. Furthermore, building on the concepts introduced in inception-RPN, two filters are employed to extract local features and feed them into two convolutional layers for the purposes of detection and recognition. Xu et al. [28] have proposed RPnet framework, for automatic license plate recognition and detection, in which the detection module utilizes ten convolutional layers to extract input image features at various levels, while the recognition module employs the maximum pooling layer of the region of interest (ROI) to extract relevant features.

In [45], a deep convolutional neural network is used to learn descriptors that capture detailed information about local ridge structures. Besides, various number of learned minutia-centred deep convolutional (MDC) features, centered around minutiae points, are aggregated into fixed-length feature vectors to enhance retrieval efficiency. Wen et al. [50] have employed pre-trained CNNs to extract features from preprocessed shoe print images, and compare the matching performance of features extracted from various layers of different pre-trained CNNs in the experiments subsection. The experimental results indicate that the 14th layer of VGG19 achieves the highest performance. These researches utilize convolutional layer features as image descriptors to achieve CSI image retrieval through feature extraction and aggregation. By gradually increasing the hierarchy, convolutional layer features can capture different of information in images, ranging from

low-level local features to high-level semantic features, providing a richer and more diverse feature representation for CSI image recognition, classification, and retrieval tasks.

3.1.2 Fully Connected Layer Feature Extraction

The output feature of the fully connected layer can be considered as an abstract and synthesized representation of the input features, capturing higher-level semantic information. Bai et al. [64] have employed VGG and ResNet networks to retrain and optimize a dataset of newly-explored images, with the objective of extracting fully connected layer features from these images. Experimental results on CIIP-CSID demonstrate that the feature extraction performance using VGG and ResNet networks far exceed that of traditional methods. Furthermore, the classification accuracy of CSI image can be further improved by changing the last pool layer to a pyramid pool layer. However, the high dimensionality of the fully connected layer features tends to result in increased computational complexity and storage requirements, potentially leading to overfitting issues. To overcome overfitting issues caused by insufficient data, Zhu et al. [3] chose the AlexNet model and extracted features of the fully connected layer, adopting an “end-to-end” approach to explore the mapping relationship between tire and tread images using the transfer learning model.

3.1.3 Fusion Feature Extraction

The image retrieval method based on fused features can improve the accuracy and robustness of image retrieval, which mainly include two types: the fusion of low-level features and deep features, and the fusion between different deep features, as shown in Fig. 6. The methods of feature fusion include weighted fusion, concatenation fusion, learning fusion, and ensemble methods. Utilizing fused features for image retrieval can provide a more comprehensive and enriched feature representation, fully leveraging multi-source information. Further, it exhibits strong noise resistance,

as well as flexibility and scalability, which significantly improve the accuracy and robustness of image retrieval.

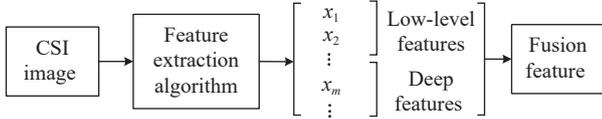


Fig. 6 The framework of feature fusion

1) Fusion of low-level features and deep features

Low-level features can capture low-level vision information in images, such as texture and color. On the other hand, deep features can capture the semantic content of images. In [51], VGG-F and VGG-VD16 were selected as pre-trained models, and chose to fuse three types of low-level features, namely HSV color histogram, Gabor features, and GIST spatial envelope features. The results demonstrated that this algorithm effectively describes the content of forensic images while maintaining a high average precision rate. In [65], transfer learning was introduced in the training of the CNN model, and a new model for tire indentation images was obtained by fine-tuning the network. The local gradient direction ternary pattern (LGDTP) feature is fused with the features extracted from the fully connected layers, producing more accurate features.

2) Fusing between deep feature

Deep neural networks have the capability to extract image features from various layers. Typically, the shallow layers of the network primarily capture low-level features, such as edges, textures, and colors. As the network layer deepens, the intermediate layers gradually uncover more complex features, such as shapes, object and structures. The deeper layers of the network can capture highly abstract and semantic features. Liu et al. [52, 66] have employed transfer learning to obtain a pre-trained CNN model, extracted features from both the convolutional and fully-connected layers of the network. Additionally, they introduced an adaptive feature learning network model, fine-tuning the parameters learned by the autoen-

coder and softmax parameters collectively. Experimental results on CIIP-CSID demonstrated that the proposed model enhances classification accuracy. Ma et al. [46] have proposed a multi-part weighted convolutional neural network (MP-CNN) for analyzing shoeprint images, which divides the shoeprint image into two parts and extracts sub-features from each part. The importance weight matrix of the sub-features is computed based on the information pixels contained in them. The final feature is obtained by fusing the convolutional layer features and fully connected layer features. Experimental evaluations demonstrate that this method achieves promising results on the SPID dataset and FID300 dataset.

3.2 Deep Metric Learning Based Retrieval

Metric Learning aims to measure the similarity or distance between samples or features by learning a metric function. Its objective is to minimize the distance between samples of the same class while increasing the distance between samples of different classes. In recent years, deep learning and metric learning have been combined and proposed the concept of deep metric learning [67]. Specifically, deep metric learning is based on the principle of similarity between samples, and uses deep learning methods to learn similarity metrics between data samples.

Deep metric learning leverages the powerful expressive and nonlinear modeling capabilities of deep neural networks by learning the mapping relationship from raw data to embedded features, in which, embedded features learned from nonlinear subspace can better capture the intrinsic structure and semantic information of the data. Deep metric learning consists of three main components: input information sample, network model structure, and metric loss function [68]. This section summarizes the CSI image retrieval methods from the perspective of network model structure, focusing on Siamese networks and Triplet networks combined with metric loss func-

tions.

3.2.1 Siamese Network

Siamese networks learn from a discriminative learning framework based on energy models. The network takes two images as input and learns to produce binary values. A value of “0” indicates that the images belong to the same class, whereas a value of “1” suggests different classes. The Siamese network, as a metric learning method, is trained by utilizing pairs of images containing both positive and negative samples [69]. The distance between image pairs is calculated using a loss function, in which the contrastive loss (Eq. (1)) exhibits superior performance for the Siamese network. This successful model effectively maximizes or minimizes the distance between objects, generating improved perfor-

mance. In deep metric learning, shared weights are used to extract meaningful patterns between images, as in Fig. 7 (a), possessing advantages in terms of time and memory efficiency. Additionally, there have been investigations into the combination of Siamese networks and convolutional neural networks in prior studies [49, 53, 70]. The contrastive loss function L_c for the Siamese network model:

$$L_c = (1 - \Gamma) \frac{1}{2} (D_w)^2 + (\Gamma) \frac{1}{2} \{\max(0, m - D_w)\}^2 \quad (3)$$

where Γ denotes the label value, taking the value of 1 when the pair of inputs belongs to the same class, and 0 otherwise. D_w is employed within the loss function to determine the distance between two inputs. The parameter m represents the margin value in L_c .

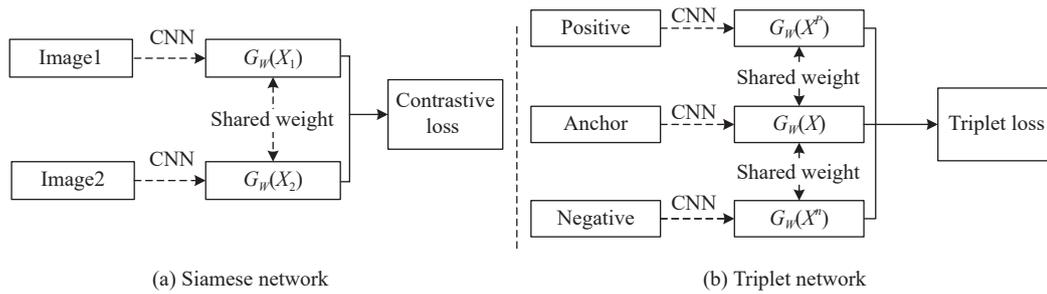


Fig. 7 Neural network models for metric learning: (a) Siamese network; (b) Triplet network

Kong et al. [53] have proposed using the multi-channel normalized cross-correlation (MCNCC) method to match features. Later, the combination of MCNCC and CCA provides an efficient building block for constructing Siamese network models, which can be trained in an end-to-end discriminative learning framework. The experiments demonstrate that even with very limited data, this framework can achieve robust cross-domain matching using a universal feature extractor combined with a simple linear feature transformation layer trained in segments, which provides state-of-the-art performance for retrieving shoe patterns matching crime scene evidence. In [71], a method based on tattoo sketches is proposed, which involves fine-tuning a CNN to obtain a model for classifying both the sketch

and the real tattoo, then uses Siamese network to train and generate feature vectors. The network outputs a similarity metric between two feature vectors, which are used for both the sketch and the real tattoo images.

3.2.2 Triplet Network

The Triplet network inspired by the Siamese network, consists of three elements: a positive sample, a negative sample, and an anchor sample [72]. In pattern recognition, Triplet networks utilize the Euclidean space to compare these elements, which directly relates to metric learning. Eq. (2) demonstrates that triplet losses prioritize the similarity between pairs of samples from the same class and different classes by utilizing shared weights. Classification is then performed by comparing the similarities of sample pairs

(Fig. 7 (b)). The utilization of Triplet networks enhances discrimination between intra-class and inter-class relationships.

The Triplet network model has three inputs: the anchor input X , the positive sample X^P , and the negative sample X^n . Triplet loss L_T :

$$L_T = \max(0, \|G_w(X) - G_w(X^P)\|_2 - \|G_w(X) - G_w(X^n)\|_2 + \alpha)^{\leftrightarrow} \quad (4)$$

where α is the margin value. In [46], footprints are divided into different regions, MP-CNN utilizes a Siamese network as the underlying structure and employs a triplet loss function as the similarity metric. Experiments were conducted on the FID-300 dataset, and the top 10% accuracy was approximately 89.8%. In [54], the TripletGAN approach is proposed to identify images of human bodies with tattoos that uses a limited number of labeled data samples. The combination of AlexNet and Siamese networks based on triplet loss, has also been employed in the retrieval of prospecting images [49, 70]. The experimental results demonstrate that the triplet loss function can enhance the performance of tattoo matching systems compared to simple contrast loss functions.

3.3 Generative Adversarial Networks Based Retrieval

A generative adversarial network (GAN) consists of a generator network and a discriminator network, in which the generator network generates new samples from random vectors in the training set, while the discriminator network distinguishes between generated images and original images. This approach can be used to increase data samples, expand datasets, or generate more similar images.

In [54], the TripletGAN model integrates the discriminator into the learning task of the triplet network, which achieved better results (0.9178) than a single Triplet network (0.9462) and a single GAN network (0.8352) on the MNIST dataset. In [73], the InfoGAN model is utilized for the retrieval of face sketches to real images.

InfoGAN is an extension of GAN that introduces control variables such as style, thickness, and type. It consists of three sub-models: a generator model, a discriminator model, and an auxiliary model for predicting the control variables, in which, the auxiliary model is constructed by including the control variables and trained using the same weights as the discriminator model, utilizing an information loss function. After the training, the discriminator can be utilized to search for visually similar images. The network provides a comprehensive and well-defined feature representation for images and estimates image similarity using spatial distance metrics derived from the extracted features.

3.4 Autoencoder Networks Based Retrieval

Autoencoder (AE) is an unsupervised neural network utilized for image reconstruction from latent spaces, consisting of an encoder and a decoder. The encoder compresses the input data into a lower-dimensional representation known as the hidden layer, which involves data compression and extraction of essential features. The decoder reconstructs the hidden layer back to the original input data dimensions, aiming to restore the input data as accurately as possible. The model is trained by minimizing the reconstruction error between the original image and the reconstructed image using a loss function.

Cao et al. [44] have proposed an end-to-end latent fingerprint search system, which includes automatic region of interest (ROI) cropping, latent image preprocessing, feature extraction, feature comparison, and outputting a candidate list. They introduced a detail detection method based on convolutional autoencoder. Sabry et al. [73] have proposed a facial sketch to real image retrieval system based on convolutional autoencoder, in which, the latent vectors that are used as image feature descriptors are obtained through the hidden layers of the encoder. The decoder utilizes the feature descriptors provided by the encoder to reconstruct the images. Then, these

attributes are combined with nearest neighbor algorithms to determine comparable and similar images. The advantage of this model is its ability to effectively represent image features in low-dimensional vectors.

3.5 Attention Networks Based Retrieval

The attention mechanism is used to focus the model’s attention on crucial parts or relevant information within the input data. In CNN models, all input features are treated uniformly without a distinct mechanism to handle interdependencies among different inputs. The attention mechanism assigns different weights to different inputs, allowing the model to selectively attend to important input information and improve the performance of the model.

In [30], the attention-based partial fingerprint recognition model (APFI) is proposed, employing ResNet to extract feature descriptors, and inserts channel attention modules into the model to obtain more accurate fingerprint feature information from the residuals. To enhance the recognition accuracy of partial fingerprints, the angle distance between features is used to calculate the similarity of fingerprints. In [55], the attention pooling method utilizes multiple functions (standard deviation, entropy, edge detection, skin segmentation algorithm) to weight the local features of the convolutional feature maps. The weighted functions emphasize the importance of local regions in tattoo images, allowing the recognition process to consider specific domain features. The best convolutional layer features and weights were selected through experiments.

Ref.[73] have proposed an image retrieval system based on vision transformer (ViT), which is trained on both facial sketches and real images. The descriptor of the image is represented by the features learned by ViT itself, and the shifted patch tokenization (SPT) is used to tokenize, and then, encode the patches, achieving local self-attention (LSA). Retrieval is performed

based on the spatial distance between features.

In [56], a novel dual knowledge distillation (DKD) network is proposed for efficient retrieval of crime scene investigation shoeprint images (SPI), which consists of a sophisticated teacher model and a lightweight student model. By using the designed distillation loss function, the spatial attention module (SA) and semantic knowledge from the teacher model are transferred to the student model to enhance its feature extraction capability and improve the accuracy of SPI retrieval.

In [57], an end-to-end deep hashing framework is proposed for fast shoeprint image retrieval and ranking, which embeds the spatial transformer network (STN) into the deep hashing network to enhance the robustness. With the aid of the triple cross-entropy loss function, the model can simultaneously achieve optimal class separability and hash code separability during the training phase. Furthermore, by utilizing small samples and triple-label learning, this approach alleviates the issue of sample imbalance to some extent.

4 Discussion

In recent years, as public safety has become a major concern, the development of CBIR technology has gained significant attention. In light of this, and considering the specific needs of public security and criminal investigation, this section suggests the following future research directions for CSIR.

4.1 Benchmark Database Construction

The existing public datasets have limited coverage in certain domains and scenarios, which hinders the generalization ability of image classification and retrieval models. Additionally, these models often perform poorly when facing new data and situations. To address this issue, it is crucial to establish a comprehensive benchmark database that encompasses various current images. This database can be used to evaluate

the performance of the models.

4.2 End-to-End System

Future CSI image retrieval systems will emphasize an end-to-end design approach, which requires the development of a comprehensive system framework that includes image acquisition, processing, storage, retrieval, and representation. By optimizing and integrating multiple components, the overall performance of the retrieval system is enhanced.

4.3 Multi-Modal Data Fusion

Image datasets are often accompanied by public security records, text, voice, and other field investigation data. By effectively integrating and associating multi-modal data, including the addition of text and voice descriptions, more comprehensive and accurate information can be provided. This integration enhances the accuracy and efficiency of current image classification and retrieval. Moreover, by combining multiple data types, the user's query intent can be expanded, resulting in more precise and comprehensive search results.

5 Conclusion

This paper begins by providing an overview of existing literature on insight image datasets, focusing on single or comprehensive types. Then, we analyze and summarize the recent advancements in deep learning-based insight image retrieval technology. These advancements can be categorized into five main areas: feature representation, metric learning, generative adversarial networks, autoencoder networks, and attention networks. Furthermore, considering the current development status and practical application requirements of the image retrieval field, we outline future research directions in this area.

References:

- [1] Y. Liu, D. Hu, and J. L. Fan, "A survey of crime scene investigation image retrieval," *Acta Electronica Sinica*, vol. 46, no. 3, pp. 761-768, 2018.
- [2] Y. Liu, Y. Ge, F. Wang, Q. Liu, Y. Lei, D. Zhang, and G. Lu, "A rotation invariant hog descriptor for tire pattern image classification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2412-2416, 2019.
- [3] J. Zhu and S. Wang, "Application of machine vision in tire and impression images classification," *Journal of Chinese Computer Systems*, vol. 42, no. 9, pp. 1967-1972, 2021.
- [4] Z. Li, C. Wei, Y. Li, and T. Sun, "Research of shoeprint image stream retrieval algorithm with scale-invariance feature transform," in *2011 International Conference on Multimedia Technology*, pp. 5488-5491, 2011.
- [5] N. Richetelli, M. C. Lee, C. A. Lasky, M. E. Gump, and J. A. Speir, "Classification of footwear outsole patterns using fourier transform and local interest points," *Forensic Science International*, vol. 275, pp. 102-109, 2017.
- [6] X. Li, M. Wu, and Z. Shi, "The retrieval of shoeprint images based on the integral histogram of the gabor transform domain," in *Intelligent Information Processing VII: 8th IFIP TC 12 International Conference, IIP 2014, Hangzhou, China, October 17-20, 2014, Proceedings 8*, pp. 249-258, 2014.
- [7] A. Kortylewski, T. Albrecht, and T. Vetter, "Unsupervised footwear impression analysis and retrieval from crime scene data," in *Computer Vision-ACCV 2014 Workshops: Singapore, Nov. 1-2, 2014, Revised Selected Papers, Part I 12*, pp. 644-658, 2015.
- [8] Y. Wu, X. Wang, and T. Zhang, "Crime scene shoeprint retrieval using hybrid features and neighboring images," *Information*, vol. 10, no. 2, pp. 45-60, 2019.
- [9] Y. Wu, X. Wang, N. L. Nankabirwa, and T. Zhang, "Logsr: learned opinion score guided shoeprint retrieval," *IEEE Access*, vol. 7, pp. 55073-55089, 2019.
- [10] X. Wang, C. Zhang, Y. Wu, and Y. Shu, "A manifold ranking based method using hybrid features for crime scene shoeprint retrieval," *Multimedia Tools and Applications*, vol. 76, pp. 21629-21649, 2017.
- [11] S. Alizadeh, H. B. Jond, V. V. Nabiyevev, and C. Kose, "Automatic retrieval of shoeprints using modified multi-block local binary pattern," *Symmetry*, vol. 13, no. 2, pp. 296-316, 2021.

- [12] A. K. Jain, J. E. Lee, R. Jin, and N. Gregg, "Content-based image retrieval: An application to tattoo images," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 2745-2748, 2009.
- [13] J. Lee, R. Jin, A. Jain, and W. Tong, "Image retrieval in forensics: tattoo image database application," *IEEE MultiMedia*, vol. 19, no. 1, pp. 40-49, 2011.
- [14] A. K. Jain, J. E. Lee, and R. Jin, "Tattooid: Automatic tattoo image retrieval for suspect and victim identification," in *Advances in Multimedia Information Processing-PCM 2007: 8th Pacific Rim Conference on Multimedia, Hong Kong, China, Dec. 11-14, 2007. Proceedings 8*, pp. 256-265, 2007.
- [15] J. E. Lee, R. Jin, and A. K. Jain, "Rankbased distance metric learning: An application to image retrieval," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2008.
- [16] S. T. Acton and A. Rossi, "Matching and retrieval of tattoo images: Active contour cbir and glocal image features," in *2008 IEEE Southwest Symposium on Image Analysis and Interpretation*, pp. 21-24, 2008.
- [17] R. Azad, B. Azad, and H. R. Shayegh, "Real-time and efficient method for accuracy enhancement of edge based license plate recognition system," *arXiv preprint arXiv: 1407.6498*, 2014.
- [18] V. Du Mai, D. Miao, R. Wang, and H. Zhang, "An improved method for vietnam license plate location, segmentation and recognition," in *2011 International Conference on Computational and Information Sciences*, pp. 212-215, 2011.
- [19] A. H. Ashtari, M. J. Nordin, and M. Fathy, "An iranian license plate recognition system based on color features," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 4, pp. 1690-1705, 2014.
- [20] H. Su and A. W. K. Kong, "A study on low resolution androgenic hair patterns for criminal and victim identification," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 4, pp. 666-680, 2014.
- [21] Y. Liu, H. Yan, and K. P. Lim, "Study on rotation-invariant texture feature extraction for tire pattern retrieval," *Multidimensional Systems and Signal Processing*, vol. 28, pp. 757-770, 2017.
- [22] Y. Liu, S. Zhang, F. Wang, K. Lim, Q. Liu, Y. Lei, Y. Gong, and J. Lu, "Waveletenergy-weighted local binary pattern analysis for tire tread pattern classification," in *2019 3rd International conference on imaging, signal processing and communication (ICISPC)*, pp. 90-95, 2019.
- [23] Y. Liu, D. Hu, J. L. Fan, F. P. Wang, and D. X. Li, "Multi-feature fusion based retrieval results optimization for crime scene investigation image retrieval," *Acta Electronica Sinica*, vol. 47, no. 2, pp. 296-301, 2019.
- [24] H. Han, J. Li, A. K. Jain, S. Shan, and X. Chen, "Tattoo image search at scale: Joint detection and compact representation learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 10, pp. 2333-2348, 2019.
- [25] Q. Xu, S. Ghosh, X. Xu, Y. Huang, and A. W. K. Kong, "Tattoo detection based on cnn and remarks on the nist database," in *2016 International Conference on Biometrics (ICB)*, pp. 1-7, 2016.
- [26] M. Nicolás-Díaz, A. Morales-González, and H. Méndez-Vázquez, "Deep generic features for tattoo identification," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 24th Iberoamerican Congress, CIARP 2019, Havana, Cuba, October 28-31, 2019, Proceedings 24*, pp. 272-282, 2019.
- [27] T. Hrkac, K. Brkic, and Z. Kalafatic, "Tattoo detection for soft biometric deidentification based on convolutional neural networks," in *Proc. OAGM-ARW Joint Workshop*, pp. 131-138, 2016.
- [28] Z. Xu, W. Yang, A. Meng, N. Lu, H. Huang, C. Ying, and L. Huang, "Towards end-to-end license plate detection and recognition: A large dataset and baseline," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 255-271, 2018.
- [29] R. Laroca, E. Severo, L. A. Zanlorensi, L. S. Oliveira, G. R. Gonçalves, W. R. Schwartz, and D. Menotti, "A robust realtime automatic license plate recognition based on the yolo detector," in *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-10, 2018.
- [30] Y. Sun, Y. Tang, and X. Chen, "A neural network-based partial fingerprint image identification method for crime scenes," *Applied Sciences*, vol. 13, no. 2, pp. 1188-1202, 2023.
- [31] G. R. Gonçalves, S. P. G. da Silva, D. Menotti, and W. R. Schwartz, "Benchmark for license plate character segmentation," *Journal of Electronic Imaging*, vol. 25, no. 5, pp. 053034-053034, 2016.
- [32] C. N. E. Anagnostopoulos, I. E. Anagnostopoulos, I. D. Psoroulas, V. Loumos, and E. Kayafas,

- “License plate recognition from still images and video sequences: A survey,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 3, pp. 377-391, 2008.
- [33] G. S. Hsu, J. C. Chen, and Y. Z. Chung, “Application-oriented license plate recognition,” *IEEE Transactions on Vehicular Technology*, vol. 62, no. 2, pp. 552-561, 2012.
- [34] I. Štajduhar and S. Sovilj, “Zemris license plate dataset,” [Online]. Available: <https://www.zemris.fer.hr/projects/LicensePlates/hrvatski/rezultati.shtml>, 2014.
- [35] J. Špaňhel, J. Sochor, R. Juránek, A. Herout, L. Maršík, and P. Zemčík, “Holistic recognition of low quality license plates by cnn using track annotated data,” in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1-6, 2017.
- [36] R. Laroca, E. V. Cardoso, D. R. Lucio, V. Estevam, and D. Menotti, “On the cross-dataset generalization in license plate recognition,” *arXiv preprint arXiv: 2201.00267*, 2022.
- [37] W. Zhou, H. Li, Y. Lu, and Q. Tian, “Principal visual word discovery for automatic license plate detection,” *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 4269-4279, 2012.
- [38] A. Kortylewski and T. Vetter, “Probabilistic compositional active basis models for robust pattern recognition,” in *British Machine Vision Conference*, 2016.
- [39] M. Ngan and P. Grother, “Tattoo recognition technology-challenge (tatt-c): An open tattoo database for developing tattoo recognition research,” in *IEEE International Conference on Identity, Security and Behavior Analysis (ISBA 2015)*, pp. 1-6, 2015.
- [40] Y. I. Shehu, A. Ruiz-Garcia, V. Palade, and A. James, “Sokoto coventry fingerprint dataset,” *arXiv preprint arXiv: 1807.10609*, 2018.
- [41] G. Fiumara, P. Flanagan, J. Grantham, B. Bandin, K. Ko, and J. Libert, “National institute of standards and technology special database 300: uncompressed plain and rolled images from fingerprint cards,” *National Institute of Standards and Technology*, Technical Note 1993, Jun. 2018.
- [42] G. Fiumara, P. Flanagan, M. Schwarz, E. Tabassi, and C. Boehnen, “National institute of standards and technology special database 301: Nail to nail fingerprint challenge dry run,” *National Institute of Standards and Technology*, Technical Note 2002, Jul. 2018.
- [43] G. Fiumara, P. Flanagan, J. Grantham, K. Ko, K. Marshall, M. Schwarz, E. Tabassi, B. Woodgate, and C. Boehnen, “Nist special database 302: Nail to nail fingerprint challenge,” *National Institute of Standards and Technology*, Technical Note 2019-12-11, 2019.
- [44] K. Cao, D. L. Nguyen, C. Tymoszek, and A. K. Jain, “End-to-end latent fingerprint search,” *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 880-894, 2019.
- [45] S. Yoon and A. K. Jain, “Longitudinal study of fingerprint recognition,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 28, pp. 8555-8560, 2015.
- [46] Z. Ma, Y. Ding, S. Wen, J. Xie, Y. Jin, Z. Si, and H. Wang, “Shoe-print image retrieval with multi-part weighted cnn,” *IEEE Access*, vol. 7, pp. 59728-59736, 2019.
- [47] X. Li, M. Wu, and Z. Shi, “Shoeprint image retrieval based on integral histogram in gabor transform domain,” *Computer Application and Software*, vol. 32, pp. 215-219, 2015.
- [48] H. Li, P. Wang, and C. Shen, “Toward end-to-end car license plate detection and recognition with deep neural networks,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 1126-1136, 2018.
- [49] X. Di and V. M. Patel, “Deep tattoo recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 51-58, 2016.
- [50] Z. Wen, J. Curran, and G. Wevers, “Shoeprint image retrieval and crime scene shoeprint image linking by using convolutional neural network and normalized cross correlation,” *Science & Justice*, vol. 63, no. 4, pp. 439-450, 2023.
- [51] Y. Liu, Y. Peng, D. Li, J. Fan, and Y. Li, “Crime scene investigation image retrieval with fusion cnn features based on transfer learning,” in *Proceedings of the 3rd International Conference on Multimedia and Image Processing*, pp. 68-72, 2018.
- [52] Q. Zhang, Y. Liu, F. Wang, J. Lu, and D. Li, “Fusion cnn based on feature selection for crime scene investigation image classification,” in *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery: Volume 2*, pp. 773-780, 2020.
- [53] B. Kong, J. Supancic, D. Ramanan, and C. C. Fowlkes, “Cross-domain image matching with deep

- feature maps,” *International Journal of Computer Vision*, vol. 127, pp. 1738-1750, 2019.
- [54] C. Jiawang and Z. Yuan, “Tattoo recognition based on triplet gan,” in *2018 37th Chinese Control Conference (CCC)*, pp. 9595-9597, 2018.
- [55] M. Nicolás-Díaz, A. Morales-González, and H. Méndez-Vázquez, “Weighted average pooling of deep features for tattoo identification,” *Multimedia Tools and Applications*, vol. 81, no. 18, pp. 25853-25875, 2022.
- [56] D. Li, Y. Li, and Y. Liu, “Shoeprint image retrieval based on dual knowledge distillation for public security internet of things,” *IEEE Internet of Things Journal*, vol. 9, no. 19, pp. 18829-18838, 2022.
- [57] W. Liu and D. Xu, “Robust and efficient shoe print image retrieval using spatial transformer network and deep hashing,” in *Proceedings of the 4th International Symposium on Signal Processing Systems*, pp. 89-95, 2022.
- [58] K. Simonyan and A. Zisserman, “Very deep convolutional networks for largescale image recognition,” *arXiv preprint arXiv: 1409.1556*, 2014.
- [59] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.
- [60] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9, 2015.
- [61] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818-2826, 2016.
- [62] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, pp.4278-4284, 2017.
- [63] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097-1105, 2012.
- [64] X. Bai, Y. Liu, and Y. Shen, “Application of deep learning in image classification of crime scene investigation,” *Journal of Xi’an University of Posts and Telecommunications*, vol. 5, pp. 43-47, 2018.
- [65] Y. Liu, H. T. Dong, F. P. Wang, and D. X. Li, “Tire indentation mark feature based on local gradient directional ternary pattern and CNN,” *Computer Simulation*, vol. 38, no. 2, pp. 399-405, 2021.
- [66] L. Ying, Z. Qian Nan, W. Fu Ping, C. Tuan Kiang, L. Keng Pang, Z. Heng Chang, C. Lu, L. G. Jun, and L. Nam, “Adaptive weights learning in cnn feature fusion for crime scene investigation image classification,” *Connection Science*, vol. 33, no. 3, pp. 719-734, 2021.
- [67] J. Lu, J. Hu, and J. Zhou, “Deep metric learning for visual understanding: An overview of recent advances,” *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 76-84, 2017.
- [68] M. Kaya and H. Ş. Bilge, “Deep metric learning: A survey,” *Symmetry*, vol. 11, no. 9, pp. 1066-1092, 2019.
- [69] I. Filković, Z. Kalafatić, and T. Hrkać, “Deep metric learning for person reidentification and de-identification,” in *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1360-1364, 2016.
- [70] X. Di and V. M. Patel, *Deep Learning for Tattoo Recognition*. Cham: Springer International Publishing, 2017, pp. 241-256. [Online]. Available: https://doi.org/10.1007/978-3-319-61657-5_10.
- [71] B. C. S. Berno, “Sketch-based multimodal image retrieval using deep learning,” Master’s thesis, Universidade Tecnológica Federal do Paraná, 2021.
- [72] E. Hoffer and N. Ailon, “Deep metric learning using triplet network,” in *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*, pp. 84-92, 2015.
- [73] E. S. Sabry, S. S. Elagooz, F. E. Abd ElSamie, W. El-Shafai, N. A. El-Bahnasawy, G. M. El-Banby, A. D. Algarni, N. F. Soliman, and R. A. Ramadan, “Image retrieval using convolutional autoencoder, infogan, and vision transformer unsupervised models,” *IEEE Access*, vol. 11, pp. 20445-20477, 2023.



Ying Liu received her Ph.D. degree from Monash University, Australia in 2007. Currently, she is a full professor at Xi’an University of Posts and Telecommunications (XUPT), China. She serves as the director of Center for Image and Information Processing in XUPT. Her

research interests include image understanding and recognition, as well as multi-modal information processing.



Aodong Zhou is a master's student at School of Communications and Information Engineering, Xi'an University of Posts and Telecommunications. Her research interests include deep learning-based image retrieval technology.



Jize Xue received the M.S. degree in control engineering and Ph.D. degree in control science and engineering from Northwestern Polytechnic University, Xi'an, China, in 2017 and 2022, respectively. He is currently an associate professor with the communications and

information engineering, Xi'an University of Posts & Telecommunications in Xi'an, China. His research interests include hyperspectral image processing, tensor modeling and pattern recognition.



Zhijie Xu is currently a professor at School of Computing and Engineering, University of Huddersfield (HUD), UK. He serves as the director of Centre for Visual and Immersive Computing in HUD, and he is a honored professor of International Joint-Research Center for

Wireless Communication and Information Processing in Xi'an University of Posts and Telecommunications. His research interests include graphics and image processing.