CriSp: Leveraging Tread Depth Maps for Enhanced Crime-Scene Shoeprint Matching

Samia Shafique[1] Shu Kong[2,3,4*] Charless Fowlkes[1*]

[1]University of California, Irvine [2]University of Macau
[3]Institute of Collaborative Innovation [4]Texas A&M University
sshafiqu@uci.edu skong@um.edu.mo fowlkes@ics.uci.edu

code and dataset at https://github.com/Samia067/CriSp

Abstract. Shoeprints are a common type of evidence found at crime scenes and are used regularly in forensic investigations. However, existing methods cannot effectively employ deep learning techniques to match noisy and occluded crime-scene shoeprints to a shoe database due to a lack of training data. Moreover, all existing methods match crime-scene shoeprints to clean reference prints, yet our analysis shows matching to more informative tread depth maps yields better retrieval results. The matching task is further complicated by the necessity to identify similarities only in corresponding regions (heels, toes, etc) of prints and shoe treads. To overcome these challenges, we leverage shoe tread images from online retailers and utilize an off-the-shelf predictor to estimate depth maps and clean prints. Our method, named *CriSp*, matches crime-scene shoeprints to tread depth maps by training on this data. *CriSp* incorporates data augmentation to simulate crime-scene shoeprints, an encoder to learn spatially-aware features, and a masking module to ensure only visible regions of crime-scene prints affect retrieval results. To validate our approach, we introduce two validation sets by reprocessing existing datasets of crime-scene shoeprints and establish a benchmarking protocol for comparison. On this benchmark, *CriSp* significantly outperforms state-of-the-art methods in both automated shoeprint matching and image retrieval tailored to this task.

Keywords: shoeprint matching · image retrieval · forensics

1 Introduction

Examining the evidence found at a crime scene assists investigators in identifying suspects. Shoeprints are likely to be found at crime scenes, despite their fewer distinct identifying features than other biometric samples like blood or hair [13]. Hence, analyzing shoeprints can help criminal justice and forensics.

Examining shoeprints forensically offers insights into the class attributes and the acquired attributes of the suspect's footwear. Class attributes pertain to the general features of the shoe, e.g., brand, model, and size. Acquired attributes encompass the unique traits delivered by the shoe with wear and tear, e.g., holes, cuts, and scratches. Our focus lies in facilitating the investigation of the class attributes of shoeprints.

* Authors share senior authorship.

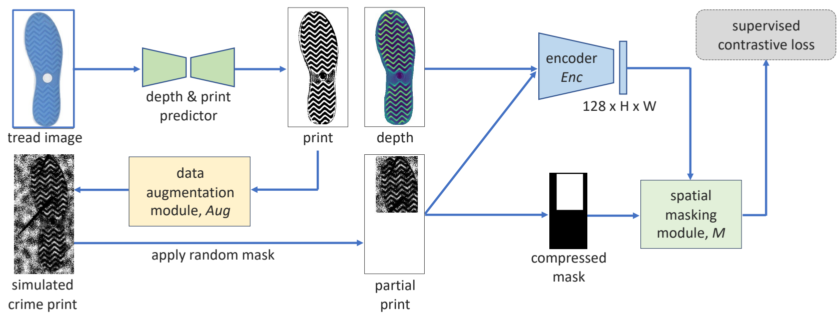

Fig. 1: Our method *CriSp* compares crime-scene shoeprints against a database of tread depth maps (predicted from tread images available at online retailers) and retrieves a ranked list of matches. We train *CriSp* using tread depth maps and clean prints (Sec. 4). We use a data augmentation module *Aug* to address the domain gap between clean and crime-scene prints, and a spatial feature masking strategy (via spatial encoder *Enc* and masking module M) to match shoeprint patterns to corresponding locations on tread depth maps (Sec. 5). *CriSp* significantly outperforms previous methods (Sec. 6).

Status quo. Traditional automated shoeprint matching methods [3, 5, 6, 14, 18, 24, 30, 31, 54] typically use handcrafted features to match crime-scene shoeprints with clean, reference impressions. Recent ones [29, 36, 60] use more generalizable features extracted by deep Convolutional Neural Networks (CNNs), which are usually pretrained on ImageNet [19] as available shoeprint datasets [32] are too small to train deep features. This solicits large-scale shoeprint datasets for better solving the shoeprint matching problem. Moreover, while existing methods match crime-scene shoeprints to clean reference shoeprints, we find that matching to tread *depth maps* leads to significantly better performance (cf. Tab. 4).

Motivation. To address the need for a large-scale training dataset, we leverage the extensive collection of tread images of various shoe products available at online retailers. We generate tread depth maps and clearly visible prints using the method propsoed in [47]. Fig. 2 shows some examples in our dataset. Note that matching directly to RGB tread images causes models to overfit to irrelevant details such as albedo and lighting (Sec. 6.3). Therefore, *we formulate our problem as the retrieval of tread depth maps that best match crime-scene shoerpints by learning a representation from tread depth maps and clean shoeprints.*

Technical insights. We develop a method termed *CriSp* to address this problem using three key components (Fig. 1). First, a data augmentation module *Aug* simulates crime-scene shoeprints from the clearly visible prints and depth maps of the training set. This helps mitigate domain gaps between our training set and real-world crime-scene testing images. Second, a spatial encoder *Enc* ensures that our model learns to match patterns in corresponding regions of shoe treads. For instance, if a crime-scene shoeprint exhibits stripes on the heels, the model must retrieve shoes with stripes in the heel region rather than other areas like the toe region. Third, a feature masking module M ensures using only the visible parts of crime-scene shoeprints for retrieval. Our extensive experiments show that combining these components facilitates feature learning and yields significantly improved retrieval performance over prior arts.

Contributions. We make three major contributions:

– We introduce the concept of matching crime-scene shoeprints to tread depth maps, aiming to facilitate forensic investigation and criminal justice.
– We propose a new benchmark consisting of a new dataset and retrieval-based evaluation protocols, allowing fair comparisons against previous methods.
– We develop a spatially-aware matching method *CriSp*, yielding superior performance over existing methods.

2 Related Work

Automated shoeprint matching. The success of automated fingerprint identification systems [17] has inspired the study of automated shoeprint matching [33, 56, 57]. Current literature aims to extract features from crime-scene shoeprints and match them to a database of laboratory footwear impressions to identify the shoe make and model [45]. Holistic methods process the shoeprint image as a whole, e.g., reconstructing the shoeprint [26], and representing shoeprints using Hu's moment [3], Zernike moment [54], and Gabor and Zernike features [30]. In contrast, local methods extract discriminative features from local regions of the shoeprint [4], making them more adept at handling partial prints. For instance, [31] exploits Wavelet-Fourier transform features, [5] introduces a block sparse representation technique, and [6] combines the Harris and the Hessian point of interest detectors with SIFT descriptors. Recent works [29, 36, 60] use features from networks pretrained on ImageNet [19]. However, the lack of large-scale shoeprint datasets hampers their effectiveness. To address this, we create a large-scale training dataset by leveraging tread images from online retailers and utilizing an off-the-shelf predictor [47] to estimate their depth maps and prints.

Image retrieval. Image retrieval techniques have been a popular research problem for several decades [61]. Traditional methods use handcrafted local features [10, 35], often coupled with approximate nearest-neighbor search methods using KD trees or vocabulary trees [11, 25, 38, 41]. More recently, the success of CNNs in classification tasks encourages their use in image retrieval tasks [9, 48]. Global features can be generated by aggregating CNN features [7, 8, 23, 39, 43, 44, 52, 53, 55], while local features can also be used for spatial verification [15, 28, 39, 41, 55] which ensure better performance by using geometric information of objects. Our problem differs from this category of work since our query and database data come from different domains - crime-scene shoeprints and depth maps of shoe treads. Even within our query set of crime-scene shoeprints, images can be from various sources such as blood, dust, and sand impressions.

Cross-domain image retrieval. More closely related to our work is cross-domain image retrieval (CDIR), where the query and database images come from different domains. The fundamental idea is to map both domains into a shared semantic feature space to alleviate the cross-domain gap. Learning a distinct representation for each shoe model can be categorized as fine-grained cross-domain image retrieval (FG-CDIR) as we aim to retrieve one instance from a gallery of same-category images. It is harder than category-level classification [20,

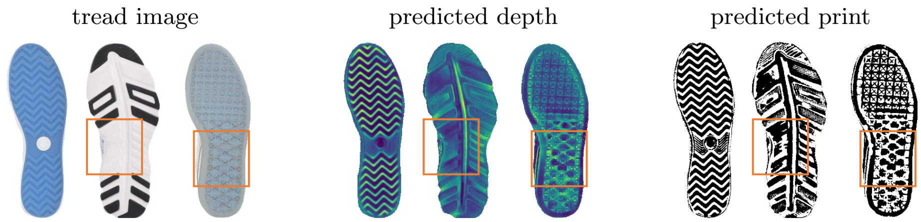

Fig. 2: Examples from train-set. We create training data from online retailers and prepare their annotations by predicting their depth maps and prints [47], although the depth and print predictions are sometimes inaccurate (2nd and 3rd shoe).

21,58] tasks since the differences between shoe treads are often subtle. A popular problem of this category, fine-grained sketch-based image retrieval (FG-SBIR), was introduced as a deep triplet-ranking based siamese network [42] for learning a joint sketch-photo manifold. FG-SBIR adopts attention-based modules with a higher order retrieval loss [50], textual tags [16, 49], and hybrid cross-domain generation [40]. The recent work [46] leverages a foundation model (CLIP) and [34] explicitly learns local visual correspondence between sketch and photo to offer explainability. These works differ from ours in that we do not have any ground-truth training data from our query domain, and thus have to simulate it as best as we can. Additionally, our aligned query and database images enable us to use spatially-aware techniques like spatial feature masking.

3 Problem Setup and Evaluation Protocol

Our goal is to retrieve shoe models that best match crime-scene impressions by comparing against a comprehensive shoe collection. We propose using tread images from online retailers to build our reference database. The problem formulation and evaluation protocol is outlined below.

3.1 Problem setup

Given an input *shoeprint* image (Fig. 4), our goal is to retrieve the most relevant *shoe tread* models from a reference database (Fig. 2). **A method should retrieve a ranked list $[r_1, r_2, ..., r_n]$ of shoe models from this database, where r_i is more likely to leave a crime-scene shoeprint similar to the input shoeprint than r_j for $i < j$.** Ranking might involve comparing learned features to represent both shoeprint images and shoe tread examples of the database. With the retrieved short-list of ranked examples, a crime-scene investigator will then examine them for further judgement.

In our work, we organize the database by storing shoe tread images and their depth maps, as prior work [47] demonstrates that using depth allows synthesizing shoeprint images as training data (cf. Sec. 4.1). Hence, we create such a database. Consequently, methods should (1) address the domain gap between crime-scene shoeprints and clean shoe tread depth maps, and (2) match partly visible shoeprints to corresponding regions of shoe tread depth maps.

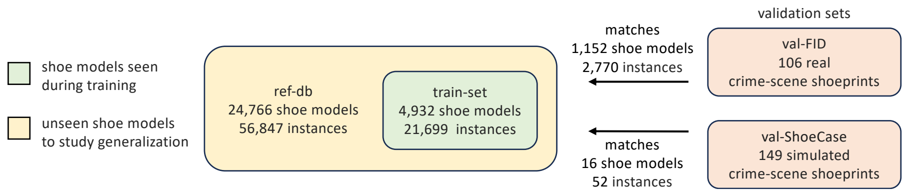

Fig. 3: Dataset statistics. We have a reference database (ref-db) and two validation sets (val-FID and val-ShoeCase) with crime-scene impressions to query against ref-db. We use a section of ref-db for training (train-set) and leave the rest to study generalization. Ground-truth labels from our validation sets connect our query crime-scene shoeprints to shoes in ref-db. See details in Sec. 4 and visual examples in Fig. 2 and 4.

3.2 Evaluation Protocol

To benchmark methods, we introduce two validation sets of crime-scene shoeprints with ground-truth shoe model labels, which are linked to a large-scale reference database (see details in Sec. 4.2). Note that the ground-truth for a shoeprint may contain multiple shoe models since tread patterns can be shared by different shoe models. In practice, we expect a human-in-the-loop approach: crime-scene investigators will look through the top K retrieved shoe models. Such a practice will greatly mitigate an open-set issue, i.e., finding that an input shoeprint does not have similar shoe models in the current database. We set K to be a realistically small value of 100, representing the top 0.4% shoe models in our reference database. We use two metrics to compare models based on their top K retrievals. Our first metric, mean average precision at K (mAP@K), is a standard metric to compare ranking performance. It considers both the number of positive matches and their positions in the ranking list. The second metric, hit ratio at K (hit@K), is more intuitive and represents the fraction of times we get at least one positive match in the top K retrievals. This metric is useful because a positive match can be used in a query expansion step to retrieve other good matches much more effectively [22]. Both metrics have values between 0 and 1, with higher numbers representing better performance. The supplement has further details.

4 Dataset Preparation

We train our model on a dataset (train-set) of aligned shoe tread depth maps and clean shoeprints. To study the effectiveness of models, we introduce a large-scale reference database (ref-db) of tread depth maps, along with two validation sets (val-FID and val-ShoeCase) created by reprocessing existing datasets of crime-scene shoeprints [32, 51]. We match shoeprints from the validation sets to ref-db and add labels connecting shoeprints in val-FID and val-ShoeCase to ref-db to enable quantitative analysis. An overview of the datasets is provided in Fig. 3, while Fig. 2 and Fig. 4 present example depth maps, clean prints, and crime-scene prints. In this section, we elaborate on our training dataset (train-set), reference database (ref-db), and validation sets (val-FID and val-ShoeCase).

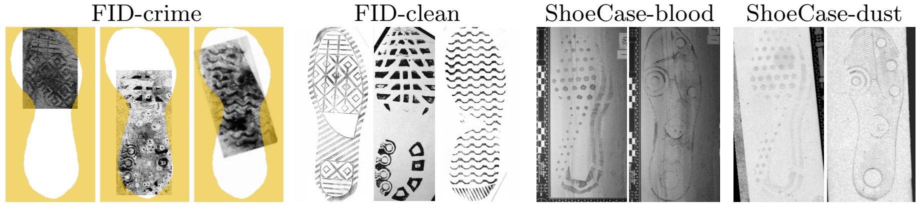

Fig. 4: Examples from val-FID and val-ShoeCase. Val-FID contains real crime-scene prints (FID-crime) and clean, fully visible lab impressions (FID-clean). We show FID-crime and FID-clean shoeprints corresponding to the same shoe models for easier comparison. Note that we show a yellow shoe outline on the FID-crime prints for visualization purposes and the outline does not exist in FID-crime images. Val-ShoeCase contains simulated crime-scene shoeprints on blood (ShoeCase-blood) and dust (ShoeCase-dust). All val-ShoeCase prints are full-sized, as opposed to val-FID.

4.1 Online Shoe Tread Depth Maps and Prints for Training

Train-set. Online retailers [1, 2] showcase images of shoe treads for advertisement. Our training set (train-set) contains depth maps and clean, fully visible prints from such tread images as predicted by [47]. We also apply segmentation masks as suggested by [47] to the predictions. To ensure consistency across all images, we employ a global alignment method to minimize variations in scale, orientation, and center using a simple model. Fig. 2 displays some sample shoe-tread images along with their corresponding depth and print predictions. Online retailers categorize shoe styles using stock keeping units (SKUs), which we use as shoe model labels. Shoes with the same SKUs can have different colors and sizes. Different shoe models may share the similar tread pattern, making them appear to be duplicates; we do not remove such likely duplicates as investigators will still examine them from the retrieved examples for the final judgement.

Statistics. Train-set contains 21,699 shoe instances from 4,932 different shoe models. Each shoe model in our database can have shoe-tread images from multiple shoe instances, possibly with variations in size, color, and lighting. The tread images in train-set have a resolution of 384×192.

Inaccuracies. It is important to note that the training dataset can have some inaccuracies since it comes from raw data downloaded from online retailers. Some tread images might have incorrect model labels, and some images may not depict shoe treads. Other inaccuracies come from imperfect depth and print prediction (cf. Fig. 2), segmentation errors, and alignment failures. We hope to mitigate the errors by including multiple instances per shoe model in train-set.

4.2 Reference Database and Crime-scene Shoeprints for Validation

Ref-db. We introduce a reference database (ref-db) by extending train-set to include more shoe models. The added shoe models are used to study generalization to unseen shoe models. Ref-db contains a total of 56,847 shoe instances from 24,766 different shoe models. The inclusion of multiple instances per shoe model in ref-db allows the depth predictor some margin for error (cf. Fig. 2),

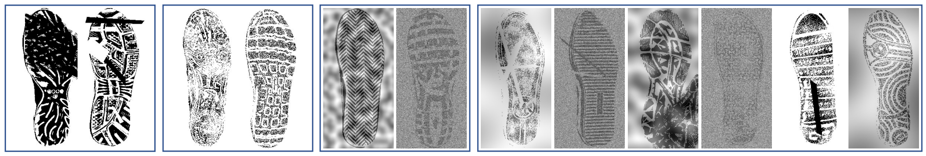

Fig. 5: Examples of data augmentation. Our data augmentation module *Aug* simulates crime-scene shoeprints (cf. Fig. 4) from clean, fully visible prints in our training set (cf. Fig. 2). *Aug* optionally (1) introduces occlusion such as overlapping prints and random shapes, (2) erases parts of the print to create a grainy appearance, and (3) adds noise to mimic background clutter.

ensuring minimal impact on the overall matching algorithm performance since it has multiple chances to match a query print to a shoe model. The supplement has details on the distribution of shoe models from our validation sets in ref-db.

Val-FID. We reprocess the widely used FID300 [32] to create our primary validation set (val-FID). Val-FID contains real crime-scene shoeprints (FID-crime) and a corresponding set of clean, fully visible lab impressions (FID-clean). Examples of these prints are shown in Fig. 4. The FID-crime prints are noisy and often only partially visible. It contains impressions made by blood, dust, etc on various kinds of surfaces including hard floors and soft sand. To ensure alignment with ref-db, we preprocess FID-crime prints by placing the partial prints in the appropriate position on a shoe "outline" (cf. Fig. 4), a common practice in shoeprint matching during crime investigations.

We manually found matches to 41 FID-clean prints in ref-db by visual inspection. These are all unique tread patterns and correspond to 106 FID-crime prints. Given that multiple shoe models in ref-db can share the same tread pattern, we store a list of target labels for each shoeprint in FID-crime. These labels correspond to 1,152 shoe models and 2,770 shoe instances in ref-db (cf. Fig. 3).

Val-ShoeCase. We introduce a second validation set (val-ShoeCase) by reprocessing ShoeCase [51] which consists of simulated crime-scene shoeprints made by blood (ShoeCase-blood) or dust (ShoeCase-dust) as shown in Fig. 4. These impressions are created by stepping on blood spatter or graphite powder and then walking on the floor. The prints in this dataset are full-sized, and we manually align them to match ref-db.

ShoeCase uses two shoe models (Adidas Seeley and Nike Zoom Winflow 4), both of which are included in ref-db. The ground-truth labels we prepare for val-ShoeCase include all shoe models in ref-db with visually similar tread patterns as these two shoe models since we do not penalize models for retrieving shoes with matching tread patterns but different shoe models. Val-ShoeCase labels correspond to 16 shoe models and 52 shoe instances in ref-db (cf. Fig. 3).

5 Methodology

In this section, we introduce *CriSp*, our representation learning framework to match crime-scene shoeprint images S to tread depth maps d. An overview of our training pipeline is shown in Fig. 1. *CriSp* is trained using a dataset of globally aligned tread depth maps d and clean, fully-visible shoeprints s (see details in Sec. 4.1). The main components of our pipeline are (1) a data augmentation module Aug that simulates crime-scene shoeprints, (2) an encoder network Enc that maps depths and shoeprints to a spatial feature representation, and (3) a spatial masking module M that masks out irrelevant portions from partially visible shoeprints.

Data augmentation. Our data augmentation module Aug simulates noisy and occluded crime-scene shoeprints (cf. Fig. 4) from clean, fully-visible prints (cf. Fig. 2), denoted as $\hat{S} = Aug(s)$. Aug uses three kinds of degradations (occlusion, erasure, and noise) as visualized in Fig. 5. Occlusion can be in the form of overlapping prints or random shapes. Erasures achieve the grainy texture of crime-scene prints and noise adds background clutter to the images. Further details are provided in the supplement.

Encoder for spatial features. Our encoder Enc maps tread depths d and simulated crime-scene shoeprints $\hat{S}$ to a feature representation z, denoted as $z = Enc(x)$ where $x \in [d, \hat{S}]$. Enc consists of a modified ResNet50 [27] with the final pooling and flattening operation removed followed by a couple of convolution layers. Enc produces features of shape $[C, H, W]$ where C is the feature length ($C = 128$ in our work), and H and W are the encoded height and width, respectively. As our training data and query prints are globally aligned (cf. Sec. 4), Enc allows access to features at each (course) spatial location of the image, facilitating comparisons in corresponding locations of shoe treads. Enc has two input channels for depth and print, respectively. It processes only one input at a time and pads the other input channel with zeros.

Spatial feature masking. During training, we simulate partially visible crime-scene shoeprints by applying a random rectangular mask m to query prints. Our feature masking module M applies a corresponding mask to spatial features z to obtain $\bar{z} = M(z, m)$. M resizes mask m to a dimension of $[H, W]$, uses it to zero out spatial features outside the mask, and normalizes the masked features. This allows our model to focus on the visible portion of the prints. While it would make sense to apply mask m to tread depth images as well, we opt not to do this as it would necessitate recomputing all the database depth features for each query print image at inference time, which is not scalable.

Training loss and similarity metric. We train our model using supervised contrastive learning [28], which extends self-supervised contrastive learning to a fully supervised setting to learn from data using labels. For a set of N depth/print pairs $\{d_k, s_k\}_{k=1...N}$ from shoe models $\{l_k\}_{k=1...N}$ within a batch, and a randomly generated mask m per batch, we compute masked spatial features $\{\bar{z}_i\}_{i=1...2N}$ and corresponding shoe labels $\{\bar{l}_i\}_{i=1...2N}$ where $\bar{z}_{2k} = M(Enc(d_k), m)$, $\bar{z}_{2k+1} = M(Enc(Aug(s_k)), m)$, and $\bar{l}_{2k} = \bar{l}_{2k+1} = l_k$. We treat $\bar{z}$ as a vector of size CHW and apply the following loss.

$$\mathcal{L} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\bar{z}_i \cdot \bar{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\bar{z}_i \cdot \bar{z}_a / \tau)} \tag{1}$$

Here, $i \in I \equiv \{1 \dots 2N\}$, $A(i) \equiv I \setminus \{i\}$, and $P(i) \equiv \{p \in A(i) : \bar{l}_p = \bar{l}_i\}$ is the set of indices of all positives in the batch distinct from i. $|P(i)|$ is the cardinality of $P(i)$. The $\cdot$ symbol denotes the inner product, and $\tau \in \mathcal{R}^+$ is a scalar temperature parameter. This loss corresponds to using cosine similarity to measure similarity between images.

Sampling. For the above loss to be effective, we must have (enough) positive examples within a batch. However, if we uniformly sample shoe models from the large-scale dataset of a large number of shoe models, a training batch might contain unique shoe models that does not have pairs of positive examples. Therefore, we sample training data in pairs, i.e. we choose $N/2$ shoe models randomly and select two random instances from each shoe model.

6 Experiments

We evaluate our *CriSp* and compare it with state-of-the-art methods on automated shoeprint matching [29] and image retrieval [28,34,46,55]. We begin with visual comparison and quantitative evaluation, followed by an ablation study and analysis of our design choices. We release our dataset and make our code publicly available at `https://github.com/Samia067/CriSp`.

6.1 Qualitative Results of CriSp

Fig. 6 shows the top 10 retrievals of our method *CriSp* on the val-FID and val-ShoeCase datasets. Notable, *CriSp* can retrieve a positive match very early even when the shoeprint has significantly limited visibility or is severely degraded. These retrievals show how *CriSp* effectively matches distinctive patterns from corresponding regions of the tread. Fig. 7 shows a comparison with related methods fine-tuned on our dataset. Clearly, *CriSp* performs significantly better at retrieving positive matches early. See more visualizations in the supplement.

6.2 Comparison with State-of-the-art

CriSp consistently outperforms previous methods across most validation examples (details in the supplement). Table 1 and 2 list comparisons on our two evaluation metrics introduced in Sec. 3.2. We analyze these results below.

Comparison with shoeprint matching. MCNCC [29] employs features from pretrained networks on ImageNet for automated shoeprint matching. However, leveraging learning on shoeprint-specific data, *CriSp* exhibits superior performance on both val-FID (see Tab. 1) and val-ShoeCase (see Tab. 2). Although MCNCC proposes to use clean shoeprint impressions as the reference database to match with, we use tread depth maps to be consistent with other methods and to achieve enhanced results. More details are in the supplement.

Comparison with image retrieval. Table 1 and 2 demonstrate how our *CriSp* consistently outperforms state-of-the-art methods in image retrieval (SupCon [28], FIRe [55], SketchLVM [46], ZSE-SBIR [34]). We fine-tune these methods on our training data containing tread depth maps and clean, fully-visible

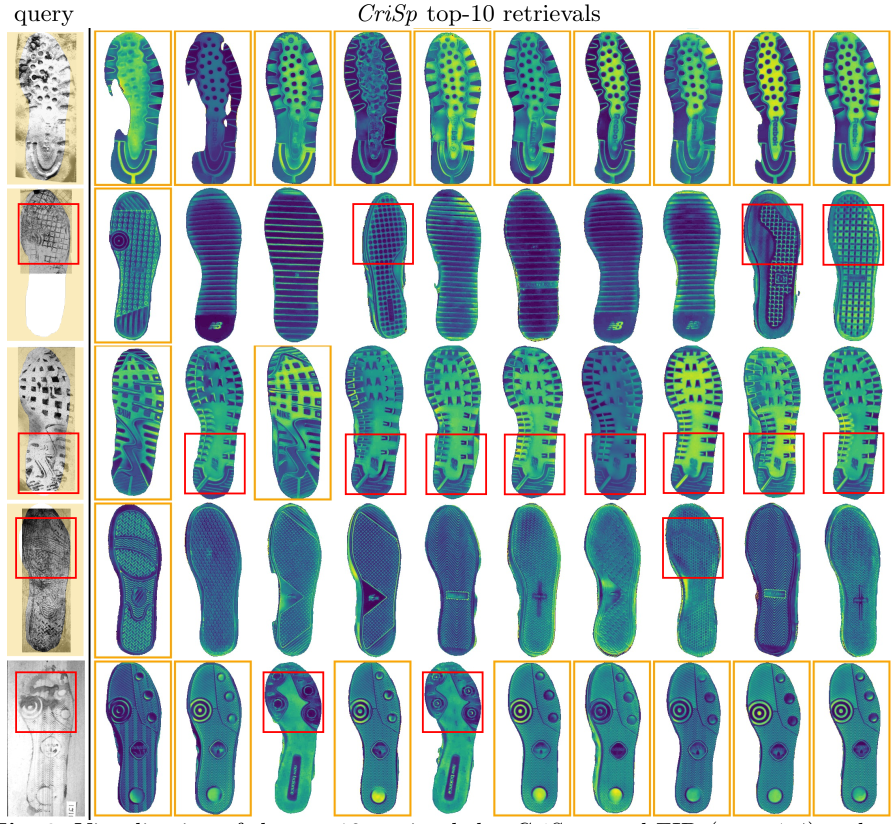

Fig. 6: Visualization of the top 10 retrievals by *CriSp* on val-FID (rows 1-4) and val-ShoeCase (row 5). *CriSp* retrieves positive matches (highlighted by orange frames) even when crime-scene shoeprints have very limited visibility or severe degradation. Additionally, corresponding locations on the retrieved shoes share similar patterns to the query print, even in negative matches (marked by red boxes).

shoeprints. Additionally, we use our data augmentation module *Aug* to simulate crime-scene shoeprints while training prior methods as the wide domain gap between crime-scene prints and the training data causes them to perform poorly otherwise (Tab. 1). Even when prior methods use our data augmentation, *CriSp* significantly outperforms them on both val-FID (Tab. 1) and val-ShoeCase (Tab. 2). The ablation study (Tab. 5) shows that our spatial feature masking technique greatly improves the performance. Qualitative comparison on both validation sets in Fig. 7 also confirm that *CriSp* is better able to match shoeprint patterns to corresponding locations on tread depth maps, thus making positive retrievals early. This is reflected by our mAP@100 values when compared to prior methods on both validation sets (Tab. 1 and 2).

Scalability. In practice, when dealing with a large reference database, scalability becomes crucial. Unlike our closest competitor ZSE-SBIR [34], which necessitates the recomputation of all database features for each query, *CriSp* offers a scalable solution. It can precompute spatial database features and effi-

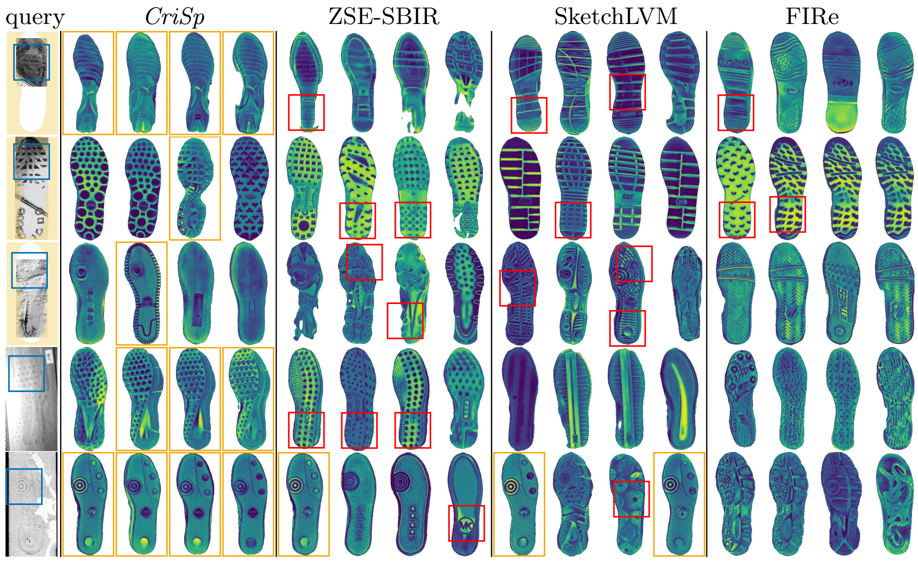

Fig. 7: Qualitative comparison with state-of-the-art methods on val-FID (rows 1-3), val-ShoeCase (rows 4-5). We show the top 4 retrieved results. *CriSp* demonstrates the ability to localize patterns, allowing it to achieve more precise retrievals (highlighted by orange frames) than previous methods. While prior methods identify similar patterns to the query print (cf. blue regions on query images), they cannot determine if they are from corresponding locations, as indicated by the red boxes in retrieved images.

ciently perform feature masking and cosine similarity calculations for each query, enabling rapid retrieval even with extensive reference databases.

Simulating partial print. Retrievals by prior methods on partial shoeprints in Fig. 7 reveal instances of poorly segmented tread depth maps, where significant portions of the tread pattern have been erased. This raises the question of whether prior methods would exhibit improved performance if trained with masks simulating partial prints. However, it is worth noting that prior methods perform better when trained without such masks, as detailed in the supplement.

Val-FID versus val-ShoeCase. Methods show a wider variation in performance on Val-ShoeCase than val-FID. This discrepancy arises from the fact that val-FID contains the diversity of real crime-scene shoeprints, while val-ShoeCase systematically simulates crime-scene prints. Additionally, val-ShoeCase contains prints from shoe models with only two unique tread patterns while val-FID contains prints from 41 unique tread patterns (cf. Sec. 4.2).

6.3 Design Choices and Ablation Study

We conduct a study of our design choices by training a ResNet50 with a supervised contrastive loss and then sequentially adding modules to investigate their performance impact. Specifically, we analyze database image configurations, data augmentation techniques, and spatial feature masking.

Table 1: Benchmarking results on real crime-scene prints from val-FID. We use hit@100 and mAP@100 as the metrics and compare previous methods trained on our dataset with / without data augmentation (cf. Sec. 5). Recall that our proposed data augmentation simulates crime-scene shoeprints from clean, fully-visible prints for the training examples. Clearly, all other prior methods benefit greatly from using our data augmentation technique. MCNCC achieves low mAP because it (1) uses off-the-shelf features from an ImageNet-pretrained model, which is not tailored to shoeprint matching, and (2) works on a more challenging and larger database (56,847 images) in our work, compared to the small-scale one (1,175 images) in its original paper [29]. SupCon also performs poorly as it samples data uniformly from the large training set that fails to guarantee enough positive pairs in training batches. Our modification (which is row-1 in Tab. 5) ensures enough positive pairs in batches through careful data sampling, yielding significant improvements. Lastly, *CriSp* significantly outperforms all the compared methods.

method		w/o our data aug		w/ our data aug	
		hit@100	mAP@100	hit@100	mAP@100
IJCV'19	MCNCC [29]	0.0849	0.0018	-	-
NeurIPS'20	SupCon [28]	0.0472	0.0020	0.0755	0.0096
ICLR'21	FIRe [55]	0.1132	0.0014	0.2075	0.0398
CVPR'23	SketchLVM [46]	0.0849	0.0066	0.1981	0.0384
CVPR'23	ZSE-SBIR [34]	0.0943	0.0065	0.4528	0.1412
	CriSp	0.0754	0.0174	**0.5472**	**0.2071**

Table 2: Benchmarking results on simulated crime-scene prints from val-ShoeCase, which includes shoeprints made by blood and dust. We use hit@100 and mAP@100 as the metrics. *CriSp* performs the best across print categories. All prior methods have been fine-tuned on our dataset using our data augmentation technique, as they perform poorly otherwise (cf. Tab. 1). Note that both ZSE-SBIR and CriSp coincidentally achieve positive matches on 62 blood prints ($62/77 = 0.8052$) and 68 dust prints ($68/72 = 0.9444$), resulting in the same hit@100, which measures the fraction of times a method gets at least one positive match within the top 100 retrievals.

method	ShoeCase-blood		ShoeCase-dust	
	hit@100	mAP@100	hit@100	mAP@100
MCNCC [29]	0.0000	0.0000	0.0000	0.0000
SupCon [28]	0.0000	0.0000	0.0000	0.0000
FIRe [55]	0.3896	0.0275	0.8194	0.3779
SketchLVM [46]	0.6623	0.1058	0.5972	0.2696
ZSE-SBIR [34]	**0.8052**	0.1849	**0.9444**	0.4063
CriSp	**0.8052**	**0.4355**	**0.9444**	**0.6792**

Database image configuration. We start by testing the effectiveness of different types of database image configurations (RGB tread images, depth, and print). Our analysis shows that depth is the most relevant and informative modality, yielding the best results when used alone (Tab. 3). Print can be derived from depth by thresholding [47] and the extra information in rgb tread images (lighting and albedo) can be distracting.

Table 3: Testing database image configurations. The hit@100 and mAP@100 values for FID-clean shoeprints indicate that using only tread depth as the database image configuration yields the best performance. Results for FID-crime are not reported in this experiment as we do not simulate crime-scene prints.

Database config.			FID-clean	
RGB	depth	print	hit@100	mAP@100
✓			0.195	0.066
	✓		**0.512**	**0.203**
		✓	0.171	0.015
✓	✓	✓	0.293	0.057

Table 4: Ablation of data augmentation techniques. We train ResNet50 networks using techniques of our data augmentation and report hit@100 and mAP@100 on FID-crime shoerpints. Results confirm that each technique (visualized in Fig. 5) individually improves retrieval results and performs best when used together.

Data augmentation			FID-crime	
occlusion	erasure	noise	hit@100	mAP@100
			0.009	0.0000
✓			0.019	0.0003
	✓		0.075	0.0098
		✓	0.170	0.0241
✓	✓	✓	**0.226**	**0.0520**

Data augmentation. Next, we test the effectiveness of each component of our data augmentation technique. Table 4 shows that all 3 components contribute to improved performance and work best when used together, bringing our hit@100 and mAP@100 on FID-crime to (0.226, 0.0520) from (0.009, 0.000).

Spatial features and feature masking. With our data augmentation in place, we study the effect of spatial feature masking, which helps *CriSp* match query print patterns to the relevant spatial locations of the database tread depth maps. Table 5 shows the influence of using spatial features and feature masking. Our findings indicate that spatial features, feature masking, and query image masking during training all contribute greatly to improving performance.

7 Discussions and Conclusions

Ethics and societal impacts. Our work is motivated by the larger goal of understanding the informational value that shoe tread pattern evidence provides in criminal investigations and forensic examination. We believe that a large dataset of tread patterns and retrieval methods will provide a positive impact as a useful resource for further studies on the human factors and uncertainty involved in making footwear-match likelihood determinations.

Court systems and footwear examiners do not generally consider matching of shoe make and model as personally identifying information (many people own the same brand of shoe) and rely on further detailed examination of acquired characteristics in conjunction with other evidence to limit false-positives. Nevertheless, there are serious broader concerns about the perils of applying artificial intelligence-based tools in the criminal justice system [37]. Similar to image retrieval in other domains, we have shown high accuracy in matching shoe tread patterns to query crime-scene evidence, but our research does not address challenging trade-offs that exist between accuracy and fairness in criminal justice risk assessments [12].

Table 5: Ablation of spatial features and feature masking. We validate the effect of using spatial features and applying feature masking on either our encoder *Enc*, which incorporates spatial features during training, or a pretrained ResNet50 which is trained with our data augmentation (cf. Tab. 4). With ResNet50 that does not utilize spatial features during training, we obtain spatial features by removing the last pooling operation. We report hit@100 and mAP@100 metrics for FID-crime shoeprints from val-FID using. Using spatial features from a pretrained ResNet50 boosts retrieval performance. Moreover, masking the spatial features improves performance further for both the ResNet50 and our *Enc*. Lastly, adding query print masking during training performs the best, yielding hit@100=0.5472 and mAP@100=0.2071.

encoder	train w/ spatial feat.	spatial features	mask features	mask query print	FID-crime	
					hit@100	mAP@100
ResNet50					0.2264	0.0520
ResNet50		✓			0.3585	0.0863
ResNet50		✓	✓		0.4245	0.1212
Enc	✓	✓			0.3774	0.1137
Enc	✓	✓	✓		0.4528	0.1765
Enc	✓	✓	✓	✓	**0.5472**	**0.2071**

We thus believe that directly applying automated shoe print retrieval methods in the real world without rigorous justification raises critical ethical issues. Ameliorating such risks in the criminal justice domain requires joint efforts from multiple communities including artificial intelligence, forensic science, criminal justice, legislative science, etc. [59]. We hope our work solicits more attention from these communities and helps foster careful application of AI-based tools (e.g., shoe print matching techniques developed in our work).

Limitations. While *CriSp* significantly outperforms prior methods on this problem, it still has some limitations. We use CNNs in our work as it is straightforward to apply the proposed spatial feature masking, yet transformer networks might perform better but it is non-trivial to mask out spatial regions in feature maps. Our work assumes that the crime-scene shoeprints are manually aligned ahead of time; methods that do not require this might be desired in the future.

Conclusion. In this paper, we propose a method to retrieve and rank the closest matches to crime-scene shoeprints from a database of shoe tread images. This is a socially important problem and helps forensic investigations. We introduce a way to learn from large-scale data and propose a spatial feature masking method to localize the search for patterns over the shoe tread. Our method consistently outperforms the state-of-the-art on both image retrieval and crime-scene shoeprint matching methods on our two validation sets that we reprocess from the widely used FID and more recent ShoeCase datasets.

Acknowledgements. This work was funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreements, 70NANB15H176 and 70NANB20H019. Shu Kong is partially supported by the University of Macau (SRG2023-00044-FST).

References

1. 6pm, http://www.6pm.com 6
2. Zappos, http://www.zappos.com 6
3. AlGarni, G., Hamiane, M.: A novel technique for automatic shoeprint image retrieval. Forensic science international **181**(1-3), 10–14 (2008) 2, 3
4. Alizadeh, S., Jond, H.B., Nabiyev, V.V., Kose, C.: Automatic retrieval of shoeprints using modified multi-block local binary pattern. Symmetry **13**(2), 296 (2021) 3
5. Alizadeh, S., Kose, C.: Automatic retrieval of shoeprint images using blocked sparse representation. Forensic science international **277**, 103–114 (2017) 2, 3
6. Almaadeed, S., Bouridane, A., Crookes, D., Nibouche, O.: Partial shoeprint retrieval using multiple point-of-interest detectors and sift descriptors. Integrated Computer-Aided Engineering **22**(1), 41–58 (2015) 2, 3
7. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5297–5307 (2016) 3
8. Babenko, A., Lempitsky, V.: Aggregating local deep features for image retrieval. In: Proceedings of the IEEE international conference on computer vision. pp. 1269–1277 (2015) 3
9. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural codes for image retrieval. In: European Conference on Computer Vision (ECCV). pp. 584–599. Springer (2014) 3
10. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer vision and image understanding **110**(3), 346–359 (2008) 3
11. Beis, J.S., Lowe, D.G.: Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (1997) 3
12. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research **50**(1), 3–44 (2021) 13
13. Bodziak, W.J.: Footwear impression evidence: detection, recovery, and examination. CRC Press (2017) 1
14. Bouridane, A., Alexander, A., Nibouche, M., Crookes, D.: Application of fractals to the detection and classification of shoeprints. In: Proceedings of International Conference on Image Processing. vol. 1, pp. 474–477 (2000) 2
15. Cao, B., Araujo, A., Sim, J.: Unifying deep local and global features for image search. In: European Conference on Computer Vision (ECCV). pp. 726–743. Springer (2020) 3
16. Chowdhury, P.N., Bhunia, A.K., Sain, A., Koley, S., Xiang, T., Song, Y.Z.: Scenetrilogy: On human scene-sketch and its complementarity with photo and text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 4
17. Datta, A.K., Lee, H.C., Ramotowski, R., Gaensslen, R.: Advances in fingerprint technology. CRC press (2001) 3
18. De Chazal, P., Flynn, J., Reilly, R.B.: Automated processing of shoeprint images based on the fourier transform for use in forensic science. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(3), 341–350 (2005) 2
19. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2009) 2, 3

20. Dey, S., Riba, P., Dutta, A., Llados, J., Song, Y.Z.: Doodle to search: Practical zero-shot sketch-based image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3
21. Dutta, A., Akata, Z.: Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5089–5098 (2019) 3
22. Efthimiadis, E.N.: Query expansion. Annual review of information science and technology (ARIST) **31**, 121–87 (1996) 5
23. Gordo, A., Almazán, J., Revaud, J., Larlus, D.: Deep image retrieval: Learning global representations for image search. In: European Conference on Computer Vision (ECCV). pp. 241–257. Springer (2016) 3
24. Gueham, M., Bouridane, A., Crookes, D.: Automatic recognition of partial shoeprints based on phase-only correlation. In: IEEE International Conference on Image Processing. vol. 4, pp. IV–441 (2007) 2
25. Han, X., Leung, T., Jia, Y., Sukthankar, R., Berg, A.C.: Matchnet: Unifying feature and metric learning for patch-based matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3279–3286 (2015) 3
26. Hassan, M., Wang, Y., Wang, D., Pang, W., Li, D., Zhou, Y., Xu, D., ur Rahman, A., Fateh, A.A., Qin, P., et al.: Deep learning model for human-intuitive shoeprint reconstruction. Expert Systems with Applications **249**, 123704 (2024) 3
27. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016) 8
28. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in Neural Information Processing Systems **33**, 18661–18673 (2020) 3, 8, 9, 12
29. Kong, B., Supancic III, J., Ramanan, D., Fowlkes, C.C.: Cross-domain image matching with deep feature maps. International Journal of Computer Vision **127**(11), 1738–1750 (2019) 2, 3, 9, 12
30. Kong, X., Yang, C., Zheng, F.: A novel method for shoeprint recognition in crime scenes. In: Biometric Recognition: 9th Chinese Conference, CCBR 2014, Shenyang, China, November 7-9, 2014. Proceedings 9. pp. 498–505. Springer (2014) 2, 3
31. Kortylewski, A., Albrecht, T., Vetter, T.: Unsupervised footwear impression analysis and retrieval from crime scene data. In: Computer Vision-ACCV 2014 Workshops: Singapore, Singapore, November 1-2, 2014, Revised Selected Papers, Part I 12. pp. 644–658. Springer (2015) 2, 3
32. Kortylewski, A., Albrecht, T., Vetter, T.: Unsupervised footwear impression analysis and retrieval from crime scene data. In: Computer Vision-ACCV 2014 Workshops: Singapore, Singapore, November 1-2, 2014, Revised Selected Papers, Part I 12. pp. 644–658. Springer (2015) 2, 5, 7
33. Li, D., Li, Y., Liu, Y.: Shoeprint image retrieval based on dual attention light hash network. In: Proceedings of the 2021 4th International Conference on Artificial Intelligence and Pattern Recognition. pp. 354–359 (2021) 3
34. Lin, F., Li, M., Li, D., Hospedales, T., Song, Y.Z., Qi, Y.: Zero-shot everything sketch-based image retrieval, and in explainable style. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23349–23358 (2023) 4, 9, 10, 12
35. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**, 91–110 (2004) 3

36. 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) 2, 3
37. Malek, M.A.: Criminal courts' artificial intelligence: the way it reinforces bias and discrimination. AI and Ethics **2**(1), 233–245 (2022) 13
38. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2, pp. 2161–2168 (2006) 3
39. Noh, H., Araujo, A., Sim, J., Weyand, T., Han, B.: Large-scale image retrieval with attentive deep local features. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 3456–3465 (2017) 3
40. Pang, K., Song, Y.Z., Xiang, T., Hospedales, T.M.: Cross-domain generative learning for fine-grained sketch-based image retrieval. In: BMVC. pp. 1–12 (2017) 4
41. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–8 (2007) 3
42. Qian, Y., Feng, L., Song, Y., Tao, X., Chen, C.L.: Sketch me that shoe. In: IEEE Conf. Comput. Vision and Pattern Recognit.(CVPR). pp. 799–807 (2016) 4
43. Radenović, F., Tolias, G., Chum, O.: Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In: European Conference on Computer Vision (ECCV). pp. 3–20 (2016) 3
44. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(7), 1655–1668 (2018) 3
45. Rida, I., Fei, L., Proen**c**ca, H., Nait-Ali, A., Hadid, A.: Forensic shoe-print identification: a brief survey. arXiv preprint arXiv:1901.01431 (2019) 3
46. Sain, A., Bhunia, A.K., Chowdhury, P.N., Koley, S., Xiang, T., Song, Y.Z.: Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2765–2775 (2023) 4, 9, 12
47. Shafique, S., Kong, B., Kong, S., Fowlkes, C.: Creating a forensic database of shoeprints from online shoe-tread photos. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 858–868 (2023) 2, 3, 4, 6, 12
48. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 806–813 (2014) 3
49. Song, J., Song, Y.Z., Xiang, T., Hospedales, T.M.: Fine-grained image retrieval: the text/sketch input dilemma. In: BMVC. vol. 2, p. 7 (2017) 4
50. Song, J., Yu, Q., Song, Y.Z., Xiang, T., Hospedales, T.M.: Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 5551–5560 (2017) 4
51. Tibben, A., McGuire, M., Renfro, S., Carriquiry, A.: Shoecase: A data set of mock crime scene footwear impressions. Data in Brief **50**, 109546 (2023) 5, 7
52. Tolias, G., Avrithis, Y., Jégou, H.: Image search with selective match kernels: aggregation across single and multiple images. International Journal of Computer Vision **116**, 247–261 (2016) 3
53. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of cnn activations. arXiv preprint arXiv:1511.05879 (2015) 3

54. Wei, C.H., Gwo, C.Y.: Alignment of core point for shoeprint analysis and retrieval. In: International Conference on Information Science, Electronics and Electrical Engineering. vol. 2, pp. 1069–1072. IEEE (2014) 2, 3
55. Weinzaepfel, P., Lucas, T., Larlus, D., Kalantidis, Y.: Learning super-features for image retrieval. In: International Conference on Learning Representations (2021) 3, 9, 12
56. Wen, Z., Curran, J., Wevers, G.: Shoeprint image retrieval and crime scene shoeprint image linking by using convolutional neural network and normalized cross correlation. Science & Justice **63**(4), 439–450 (2023) 3
57. Wu, Y., Dong, X., Shi, G., Zhang, X., Chen, C.: Crime scene shoeprint image retrieval: A review. Electronics **11**(16), 2487 (2022) 3
58. Yelamarthi, S.K., Reddy, S.K., Mishra, A., Mittal, A.: A zero-shot framework for sketch based image retrieval. In: European Conference on Computer Vision (ECCV). pp. 300–317 (2018) 3
59. Zavr**v**snik, A.: Criminal justice, artificial intelligence systems, and human rights. In: ERA forum. vol. 20, pp. 567–583. Springer (2020) 14
60. Zhang, Y., Fu, H., Dellandréa, E., Chen, L.: Adapting convolutional neural networks on the shoeprint retrieval for forensic use. In: Chinese Conference on Biometric Recognition. pp. 520–527. Springer (2017) 2, 3
61. Zheng, L., Yang, Y., Tian, Q.: Sift meets cnn: A decade survey of instance retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(5), 1224–1244 (2017) 3