Tattoo Image Search at Scale: Joint Detection and Compact Representation Learning

Hu Han, *Member, IEEE*, Jie Li, Anil K. Jain, *Fellow, IEEE*,
Shiguang Shan, *Senior Member, IEEE*, and Xilin Chen, *Fellow, IEEE*

Abstract—The explosive growth of digital images in video surveillance and social media has led to the significant need for efficient search of persons of interest in law enforcement and forensic applications. Despite tremendous progress in primary biometric traits (e.g., face and fingerprint) based person identification, a single biometric trait alone can not meet the desired recognition accuracy in forensic scenarios. Tattoos, as one of the important soft biometric traits, have been found to be valuable for assisting in person identification. However, tattoo search in a large collection of unconstrained images remains a difficult problem, and existing tattoo search methods mainly focus on matching cropped tattoos, which is different from real application scenarios. To close the gap, we propose an efficient tattoo search approach that is able to learn tattoo detection and compact representation jointly in a single convolutional neural network (CNN) via multi-task learning. While the features in the backbone network are shared by both tattoo detection and compact representation learning, individual latent layers of each sub-network optimize the shared features toward the detection and feature learning tasks, respectively. We resolve the small batch size issue inside the joint tattoo detection and compact representation learning network via random image stitch and preceding feature buffering. We evaluate the proposed tattoo search system using multiple public-domain tattoo benchmarks, and a gallery set with about 300K distracter tattoo images compiled from these datasets and images from the Internet. In addition, we also introduce a tattoo sketch dataset containing 300 tattoos for sketch-based tattoo search. Experimental results show that the proposed approach has superior performance in tattoo detection and tattoo search at scale compared to several state-of-the-art tattoo retrieval algorithms.

Index Terms—Large-scale tattoo search, joint detection and representation learning, sketch based search, multi-task learning

1 Introduction

In the past few decades, because of the advances in computing, imaging, and Internet technologies, digital images and videos are now widely used for representing information in video surveillance, and social media. In 2017, IHS Markit estimated that the United States has approximately 50 million surveillance cameras, and China has about 176 million.[1] Another statistics by YouTube in 2017 shows that the length of videos uploaded to YouTube every minute is approximately 300 hours.[2] Given the explosive growth of image and video data, there is great demand for efficient instance search technologies, particularly for persons of interest in law enforcement and forensics. Although tremendous progress has been made in face recognition based person identification, many situations exist where face recognition cannot identify an individual with sufficiently high accuracy. This is especially true when the face image quality is poor or the persons of interest intentionally hide their faces. In such cases, it is critical to acquire supplementary information to assist in person identification [1]. On the basis of this rationale, since 2009 the US Federal Bureau of Investigation (FBI) has been working on extending the capability of its Integrated Automated Fingerprint Identification System (IAFIS) by using additional biometric modalities, including iris, palm print, scars, marks, and tattoos[3]. This new person identification system is named the Next Generation Identification (NGI) system,[4] which is able

- *H. Han is with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China, and also with Peng Cheng Laboratory, Shenzhen, China, and the University of Chinese Academy of Sciences, Beijing 100049, China. E-mail: hanhu@ict.ac.cn.*
- *J. Li and X. Chen are with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China, and University of Chinese Academy of Sciences, Beijing 100049, China.*
E-mail: jie.li@vipl.ict.ac.cn, xlchen@ict.ac.cn.
- *A. K. Jain is with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824.*
E-mail: jain@cse.msu.edu.
- *S. Shan is with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China and University of Chinese Academy of Sciences, Beijing 100049, China, and he is also a member of CAS Center for Excellence in Brain Science and Intelligence Technology.*
E-mail: sgshan@ict.ac.cn.

Manuscript received 15 July 2018; revised 31 Oct. 2018; accepted 23 Dec. 2018. Date of publication 8 Jan. 2019; date of current version 12 Sept. 2019.

Recommended for acceptance by L. Liu, M. Pietkäinen, J. Chen, G. Zhao, X. Wang, and R. Chellappa.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2019.2891584

[1] http://www.straitstimes.com/opinion/chinas-all-seeing-surveillance-state-is-reading-its-citizens-faces
[2] https://www.statisticbrain.com/youtube-statistics
[3] Scars, marks, and tattoos are collectively referred to as SMT.
[4] https://www.eff.org/files/pender-fbi-next-generation-identification-overview.pdf.

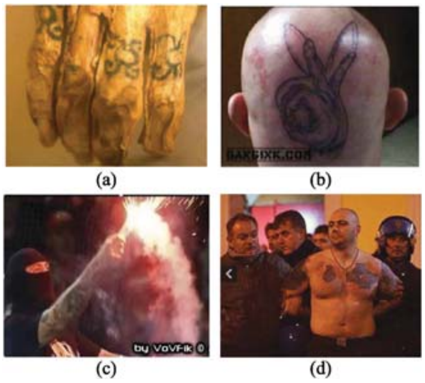


Fig. 1. Tattoo examples: (a) A tattoo on the right hand of a Chiribaya mummy in southern Peru who lived from A.D. 900 to 1350,[5] (b) a tattoo on the head signifying gang membership association,[6] (c, d) tattoos of a masked ringleader of the riots during Euro 2012 qualifier, and the suspect of the masked ringleader identified by tattoos.[7]

to offer state-of-the-art biometric identification services for homeland security, law enforcement, etc.

Among the various soft biometric traits, tattoos, in particular, have received substantial attention over the past several years due to their prevalence among the criminal section of the population and their saliency in visual attention. Humans have marked their bodies with tattoos to express personal beliefs or to signify group association for more than 5,000 years (see Fig. 1a). Figs. 1c and 1d show an example how a suspect of the masked ringleader of the riots during Euro 2012 qualifier was arrested based on the tattoos on his arms. In fact, criminal investigations have leveraged soft biometric traits as far back as the late 19th century [2]. For example, the first personal identification system, the Bertillon system, tried to provide a precise and scientific method to identify criminals by using physical measurements of body parts, especially measurements of the head and face, as well as the images of SMT on the body. Tattoos were also reported to be useful for assisting in identifying victims of terrorist attacks such as 9/11 and natural disasters like the 2004 Indian Ocean tsunami [3]. Nowadays, law enforcement agencies in the US routinely photograph and catalog tattoo patterns for use in identifying victims and convicts. The NIST Tatt-C and Tatt-E challenges have been helpful in advancing the development of tattoo detection and identification systems for real application scenarios [4].

Despite the value of tattoos for assisting in person identification, putting it to practical use has been difficult. Unlike primary biometric traits, much variability exists in pattern types of tattoos. The early use of tattoo for assisting in person identification relied heavily on manual annotations and comparisons. This has motivated the study of automatic identification algorithms for tattoos [3], [4], [5], [6], [7], [8], [9], [10].

Despite the progress in tattoo retrieval, existing methods have some serious limitations. In fact, most of the current practice of tattoo matching aims at tattoo identification, and not learning a compact representation for efficient tattoo search at scale. More importantly, existing tattoo identification methods primarily focus on matching cropped tattoos, which does not replicate the in-situ scenarios, where the search must be operated in raw images or video frames. In addition, while sketch to photo matching has been widely studied in the areas such as image retrieval [11], [12] and face recognition [13], [14], research on sketch based tattoo search is very limited [9].

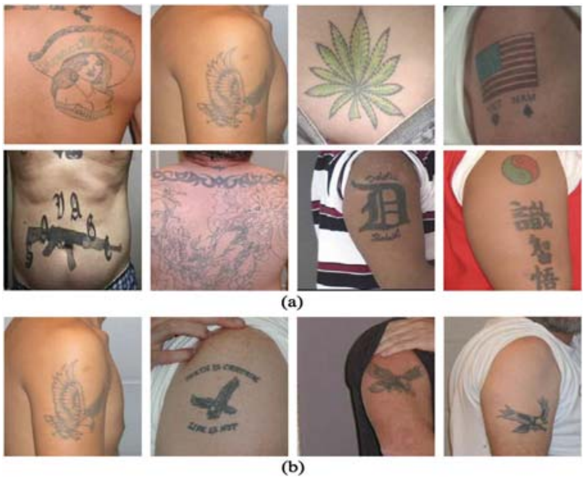


Fig. 2. Tattoo images from the Tatt-C dataset [5] suggest that tattoo search is challenging because (a) there are various types of tattoos containing categories such as humans, animals, plants, flags, objects, abstract, symbols, etc., and (b) the intra-class variability in one class of tattoos (i.e., eagle) can be very large.

1.1 Proposed Approach

To overcome the above limitations of the current tattoo search methods, we present a joint detection and compact representation learning approach for tattoo search at scale. The proposed approach is motivated by recent advances in object detection and compact representation learning, but takes into account the unique challenges in a tattoo search domain, such as large intra-class variability, poor image quality, and image deformations (see Fig. 2). In addition, the proposed approach can be trained in a fully end-to-end fashion, and can leverage additional operational data to improve the tattoo search.

As shown in Fig. 3, the proposed approach handles tattoo detection and compact representation learning in a single convolutional neural network (CNN) via multi-task learning. Given an input image, a shared feature map is first computed via a deep CNN network, which is then fed into individual sub-networks, which aim at tattoo detection and compact feature learning tasks, respectively.

The main contributions of this paper include: (i) the first end-to-end trainable approach for joint tattoo detection and compact representation learning in a single network allowing more robust and discriminative feature learning via feature sharing; (ii) effective strategies in resolving small batch size issue w.r.t. the compact representation learning module of the network; (iii) superior performance and much lower computational cost compared to the state-of-the-art algorithms; and (iv) compiling a dataset with 300,000 tattoo images in the wild and thousands of annotations for large-scale tattoo search, and a dataset with 300 tattoo sketches (see Fig. 10b) for sketch-based tattoo search; both datasets will be put into the public domain.

[5] https://www.smithsonianmag.com/history/tattoos-144038580
[6] http://www.gangink.com/index.php?pr=GANG_LIST
[7] http://www.telegraph.co.uk/sport/football/teams/serbia/8061619/Masked-ringleader-of-crowd-trouble-during-Italy-Serbia-clash-identified-by-tattoos.html

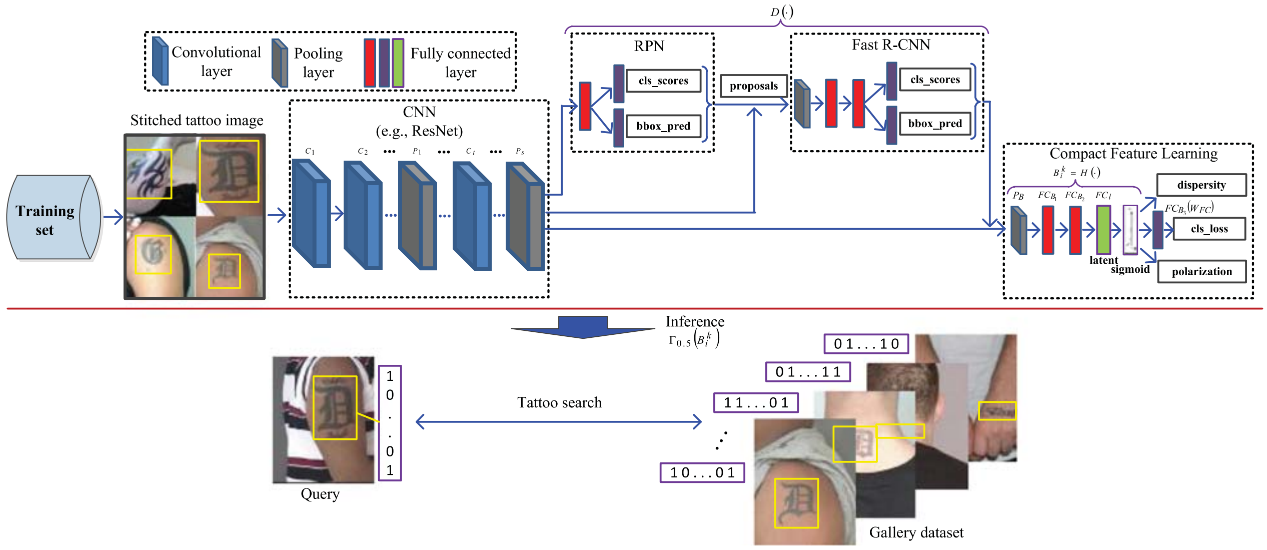


Fig. 3. Overview of the proposed approach for tattoo search at scale via joint tattoo detection and compact representation learning. Our approach consists of a stem CNN for computing the shared features, an RPN [32] and Fast R-CNN [66] for tattoo detection, and a compact representation learning module. The proposed approach can be trained end-to-end via stochastic gradient descent (SGD) [30] and back-propagation (BP) [67].

Our preliminary work of this research is described in [9]. Essential improvements over [9] include: (i) use of task-driven learned features instead of hand-crafted features for joint tattoo detection and compact representation learning in a single network; (ii) extensions to large-scale tattoo search via compact feature learning; and (iii) compiling a sketch dataset for studying sketch-based tattoo search.

The remainder of this paper is structured as follows. We briefly review related literature in Section 2. The details of the proposed tattoo search approach are provided in Section 3. In Section 4, we introduce the WebTattoo and tattoo sketch datasets, and provide the experimental results and analysis. Finally, we conclude this work in Section 5.

2 Related Work

2.1 Tattoo Identification and Retrieval

In the following, we briefly review the literature on tattoo identification and retrieval, covering detection, feature representation, databases, and performance (see Table 1).

The early practice of tattoo image retrieval relied on keywords or metadata based matching. For example, law enforcement agencies usually follow the ANSI/NIST-ITL 1-2000 standard [5] for assigning a single keyword to each tattoo image in the database. However, a keyword-based tattoo image retrieval has several limitations in practice [1]: (i) The classes defined by ANSI/NIST-ITL offer a limited vocabulary which is insufficient for describing various tattoo patterns; (ii) multiple keywords may be needed to adequately describe a tattoo image; (iii) human annotation is subjective and different subjects can give dramatically different labels to the same tattoo image.

These shortcomings of keyword-based tattoo image retrieval systems have motivated the development of content-based image retrieval (CBIR) techniques to improve the tattoo search efficiency and accuracy [1], [3], [6], [7], [8], [9], [10]. CBIR aims to extract features, e.g., edge, color, and texture, that can reflect the content of an image, and use them to identify images with high visual similarity. For example, color histogram and correlogram, shape moments, and edge direction coherence features were used in [3], [6] for tattoo matching. Similarly, global and local features of edge and color were used in [7], and vector-wise euclidean distance was computed to measure the similarity between two tattoo images. The bag-of-words (BoW) model [24], [25] using SIFT [26] features is probably the most popular one among the early CBIR systems for tattoo search [1], [8], [9], [15], [16]. Besides SIFT features, LBP-like features and HoG features were also used in [17], [18] with SVM and random forest classifiers [27] for tattoo classification. While these CBIR systems are reported to provide reasonably high accuracies on various benchmarks, they require careful handcrafting of feature descriptor, vocabulary size, and indexing algorithm.

With the success of deep learning in many computer vision tasks [28], the focus of CBIR methods is shifting from handcrafted features and models to deep learning based methods [29]. In particular, AlexNet [30], winning the ImageNet challenge of 2012, has been successfully used for tattoo versus non-tattoo classification in [10], [22], [23]. Faster R-CNN [31] was used for both tattoo versus non-tattoo image classification and tattoo localization in [21]. Among these methods [10], [21], [22], only [22], [23] studied the tattoo identification using a Siamese network with triplet loss.

There are some studies on logo or landmark search, which faces similar challenges to tattoo search, e.g., geometric deformation, inverted brightness, etc. Due to these challenges, image search algorithms based on the traditional BoW representation often fail. To resolve these issues, great efforts have been made to improve the robustness [32], [33] and efficiency [34], [35] of the descriptors. Due to space limitation, we refer interested readers to recent reviews of general image retrieval, e.g., [29].

The current practice of tattoo matching in the literature is towards tattoo identification, and not learning a compact representation for efficient large-scale tattoo search. Besides, most of the existing tattoo identification methods (including the logo or landmark search algorithms) focus on matching cropped instances (where instance of interest has been

TABLE 1
A Summary of Published Methods on Tattoo Identification and Retrieval

Publication	Detection model	Feature and retrieval model	Tattoo database #images (query; target)	Results

Jain et al. [1], [4] (2007,2012)	Gradient thresholding	Color histogram and correlogram;
		shape moments;
		edge direction coherence;
		Fusion of per feature similarities
			Tattoos from web (2,157; 43,140)[1]	46% prec.@60% recall

Acton and Rossi [8] (2008)	Active contour segmentation and skin detection	Global and local features of edge and color; Vector-wise euclidean distance	Recreational (30; $\sim 4,000$)[2]; Gang (39; $\sim 4,000$)[2]	Recreational: 94.7% acc.@rank-1; Gang: 82.2% acc. @rank-1
Jain et al. [16] (2009)	Pre-cropped tattoos	SIFT features with geometric constraint; indexing with location and keyword; Keypoint-wise matching	MSP (1,000; 63,592)	85.9% acc.@rank-1
Li et al. [17] (2009)	n/a	SIFT features; Bag-of-words; Re-ranking	MSP and ESP (995; 101,754)[3]	67% acc.@rank-1
D. Manger [9] (2012)	n/a	SIFT features; Bag-of-words, hamming embedding, and weak geometry consistency	German police (417; 327,049)	78% acc.@rank-1
Heflin et al. [18] (2012)	Automatic GrabCut and quasi connected components	LBP-like features, SVM	Tattoo classification (50; 500)[4]	85% acc.@10% FAR on average, for 15 classes
Han and Jain [10] (2013)	Pre-cropped tattoos	SIFT features; Sparse representation classification	MSU Sketch Tattoo (100; 10,100)	48% acc.@rank-100
Wilber et al. [19] (2014)	Pre-cropped tattoos	Exemplar code using HoG features; Random forest classifier	238 tattoos of 5 classes	63.8% avg. acc. for 5 classes
Xu et al. [20] (2016)	Skin segmentation and block based decision tree	Boundary features; Shape matching via coherent point drift	Full body tattoo sketch (547; 1,641)	52.38% acc.@rank-50
Kim et al. [21] (2016)	Graphcut	n/a	Tatt-C (Detection): 6,308 images; Evil (Detection): 1,105 images	Tatt-C: 70.5% acc. @41%recall Evil: 69.9% acc. @67.0%recall
Xu et al. [11] (2016)	Modified AlexNet (tattoo vs. non-tattoo)	n/a	Tatt-C (tattoo vs. non-tattoo) (1,349; 1000)[4]; Flickr (tattoo vs. non-tattoo) (5,740; 4,260)[4]	Tatt-C (tattoo vs. non-tattoo): 98.8% Flickr (tattoo vs. non-tattoo): 78.2%
Sun et al. [22] (2016)	Faster R-CNN	n/a	Tatt-C: tattoo vs. non-tattoo (1,349; 1000)[4]; Flickr: tattoo vs. non-tattoo (5,740; 4,260)[4]	Tatt-C (tattoo vs. non-tattoo): 98.25% Tatt-C (localization): 45%@0.1FPPI Flickr (tattoo vs. non-tattoo): 80.66%
Di and Patel [23], [24] (2016)	AlexNet and SVM (tattoo vs. non-tattoo)	Siamese network with triplet or contrastive loss	Tatt-C: tattoo vs. non-tattoo (1,349; 1,000)[4]; mixed media (181; 55)	Tattoo vs. non-tattoo: 99.83% Mixed media: 56.9% acc.@rank-10

TABLE 1
(*Continued*)

Publication	Detection model	Feature and retrieval model	Tattoo database #images (query; target)	Results
Proposed approach	Deep end-to-end learning for joint detection and compact representation learning		Tatt-C: detection: 7,526 images identification (157; 4,375); Flickr (detection): 5,740 images; DeMSI (identification): 890 images; WebTattoo (500, $\sim 300K$)	**Detection (localization)** Tatt-C: 61.7% recall@0.1FPPI WebTattoo: 87.1% recall@0.1FPPI **Tattoo search** WebTattoo (photo): 60.1% mAP (w/o background) 25.3% mAP (300K background) WebTattoo (sketch): 37.2% mAP (w/o background) **Tattoo identification** WebTattoo: 63.5% acc.@rank-1 (w/o background) 28.0% acc.@rank-1 (300K background) Tatt-C: 99.2% acc. @rank-1

[1] *Twenty different image transformations were applied to 2,157 tattoo images to generate 43,140 synthetic tattoo images.* [2] *Forty different image transformations were applied to 100 tattoo images to generate 40,000 synthetic tattoo images.* 3^{4}0, 000 *images were randomly selected from the ESP game dataset to populate the tattoo dataset.* [4]*(a, b) denotes the number of positive and negative tattoo images per class.* [5]*(a, b) denotes the number of tattoo and non-tattoo images.*

segmented from the background), which is different from real application scenarios, where the images are usually uncropped. Even for the deep learning based methods such as [10], [21], [22], none has addressed tattoo detection and compact representation learning jointly. The most related work of joint detection and representation learning was reported in face recognition [36], in which two large-scale face datasets, i.e., WIDER FACE [37] (containing 393,703 face bounding boxes in 32,203 images, and an average of 12 faces per image) and CASIA WebFace [38] (containing 494,414 face images of 10,575 subjects), were used to learn their end-to-end face detection and recognition network. Given the large number of faces per image on average, the small batch size is not an issue in training the recognition part of their joint detection and recognition network. In contrast, there is usually a single tattoo instance in each image in most of the tattoo datasets; this results in small batch size issue in training the recognition part of our joint tattoo detection and representation learning network because there is only one input image in each iteration. These unique challenges require design of novel end-to-end detection and representation learning approach. In addition, while we aim for efficient large-scale tattoo search and sketch-based tattoo search by performing detection and compact representation learning jointly, [36] did not report results in such a scenario.

2.2 Compact Representation Learning

Compact representation learning is of particular interest because of the need for efficient methods in large-scale visual search and instance retrieval applications [29], [39], [40], [41], [42], [43]. Compared with high-dimensional real-valued representations, compact representations aim to obtain a compressive yet discriminative feature. Feature indexing, i.e., through various quantization or hashing functions, is a major approach to obtain compact representations.

Quantization based feature indexing methods are designed to quantize the original real-valued representation with minimum quantization errors, and thus usually have high search accuracy [32], [44], [45], [46], [47], [48]. Compared with quantization based methods, hashing based feature indexing methods generates binary codes, providing faster retrieval speed since the Hamming distance of two binary codes can be computed via the native bit-wise operations. The published hashing based methods for compact representation can be categorized into two major classes: unsupervised and supervised hashing.

Unsupervised hashing algorithms, e.g., [49], [50], [51], [52], [53], [54], [55] use unlabeled data to generate binary codes which aim to preserve the similarity information in the original feature space. Unsupervised hashing methods are often efficient in computation, but their performance in large-scale retrieval may not be the optimum since no label information, including weak labels such as pairwise relationship about a dataset is utilized. To address this limitation, supervised hashing approaches, e.g., [56], [57], [58], [59], [60] have been proposed to learn more discriminative binary codes by leveraging both the label similarity and semantic similarity in the feature values of the data and label information. Deep neural networks have also been used to learn compact binary codes from high-dimensional inputs [40], [61], [62], [63]. With the advances in CNN architectures and fine-tuning strategies, the performance of the deep hashing methods is improving, and has provided good generalization ability into new datasets [30]. Due to the limited space, we refer readers to [64], [65] for a survey of data hashing approaches.

While there are a large number of approaches on hashing based compact representation learning, most of the published methods assume pre-cropped images of instances; but in a fully automatic instance retrieval system, such an assumption usually does not hold. In addition, most of the published methods on compact representation learning are designed for computer vision tasks such as face image or natural image retrieval, their performance in large-scale tattoo search is not known.

3 Proposed Method

3.1 Review of Faster R-CNN

Faster R-CNN [31] is one of the leading object detection frameworks to identify instances of objects belonging to certain classes and localize their positions (bounding boxes) in an end-to-end learning network. Faster R-CNN consists of two modules. The first module, called the Region Proposal Network (RPN), is a fully convolutional network for generating regions of interest (RoI) that denote the possible presence of objects. The second module is Fast R-CNN [66], whose purpose is to classify the RoI by RPN into individual classes and refine the positions of each foreground instance. By sharing the deep features of the full image between RPN and Fast R-CNN, Faster R-CNN is able to perform object detection accurately and efficiently, and can be trained in an end-to-end fashion.

As aforementioned, conventional methods usually break down the tattoo search problem into two separate tasks, i.e., tattoo detection [10], [21], and tattoo matching [6], [9], [18]. Such a scheme is not optimum because the matching task could assist in the detection task, and the detection accuracy influences the feature discriminability used by the matching task. Therefore, while Faster R-CNN provides an efficient solution for object detection from images, it addresses only the front-end detection problem of an instance retrieval system. Our observation is that the convolutional feature maps used by RPN and Fast R-CNN can also be used for learning compact representation.

3.2 Joint Detection and Compact Representation Learning

We aim to handle tattoo detection and compact representation learning simultaneously via a single model (see Fig. 3). A straightforward method for handling tattoo detection and compact representation learning jointly is to use a cascade of tattoo detection and compact representation learning, i.e., the output of the detector is fed into the succeeding feature extraction module. However, such a cascaded method does not leverage feature sharing to achieve efficient and robust representation learning.

Formally, let $\mathbf{A} = \{\mathbf{X}, \mathbf{Y}\}$ be a training tattoo dataset, where $\mathbf{X} = \{\mathbf{X}_i\}_{i=1}^N$ denotes N tattoo images from K distinct tattoos, and $\mathbf{Y} = \left\{\{\mathbf{Y}_i^j\}_{j=1}^M\right\}_{i=1}^N$ denotes the M labels for the corresponding tattoo images. Here, the label of each tattoo image consists of two elements (thus $M = 2$), i.e., the tattoo position ($\{\mathbf{Y}_i^1\}_{i=1}^N$) and class ID ($\{\mathbf{Y}_i^2\}_{i=1}^N$).

Given such a training dataset, we expect to jointly optimize a tattoo detector $\mathbf{D}(\cdot)$ and a compact representation learning function $\mathbf{H}(\cdot)$, which can minimize a regression loss (ℓ_{reg}) between the predicted and ground-truth bounding boxes, and a classification loss (ℓ_{cls}) of the compact representations describing individual detected tattoos, respectively. For the regression loss (ℓ_{reg}), we choose to use the robust smooth L_1 loss [66]

$$\ell_{reg}\big(\mathbf{D}(\mathbf{X}_i), Y_i^1\big) = \sum_{d\in\{u,v,w,h\}} \mathcal{S}_{L_1}\big(\mathbf{D}(\mathbf{X_i})^{\{d\}} - \mathbf{Y}_i^{1,d}\big), \qquad (1)$$

where the four-tuple $\{u, v, w, h\}$ specifies the top-left location (u, v) and the width and height (w, h) of a detected tattoo, and the four elements are indexed by d. The function $\mathcal{S}_{L_1}(\cdot)$ is defined as

$$\mathcal{S}_{L_1}(z) = \begin{cases} 0.5z^2 & \text{if } |z| < 1 \\ |z| - 0.5 & \text{otherwise} \end{cases}. \qquad (2)$$

For the classification loss (ℓ_{cls}), defined w.r.t. the detected tattoos, we choose to use the cross-entropy loss [68]

$$\ell_{cls} = -\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbf{1}\Big(\hat{\mathrm{Y}}_i^k, Y_i^2\Big)\log p\Big(\hat{\mathrm{Y}}_i^k\Big), \qquad (3)$$

where $\hat{\mathbf{Y}}_i^k = W_{FC}\cdot\mathbf{B}_i^k = W_{FC}\cdot\mathbf{H}(X_i, \mathbf{D}(X_i))^{\{k\}}$ denoting the kth element of the output by a fully connected layer with weight W_{FC}, which takes feature $\mathbf{B}_i^k$ as its input. $\mathbf{H}(\cdot)$ takes image X_i and the detected tattoo location $\mathbf{D}(X_i)$ as input, and outputs $\mathbf{B}_i^k$. $\mathbf{1}\Big(\hat{\mathbf{Y}}_i^k, Y_i^2\Big)$ outputs 1 when $k = Y_i^2$, and 0 otherwise. The probability $p(\cdot)$ is computed as

$$p(\hat{\mathrm{Y}}_i^k) = \frac{e^{\hat{\mathrm{Y}}_i^k}}{\sum_{k=1}^K e^{\hat{\mathrm{Y}}_i^k}}. \qquad (4)$$

By minimizing the losses given in (1) and (3), we can jointly perform tattoo detection and feature representation learning from the detected tattoos. However, additional constraints are still required to guarantee that the learned features are compact binary features, which is important for efficient large-scale search. Therefore, we expect that the features, i.e., $\mathbf{B}_i^k = \{b_k | b_k \in \{0,1\}\}_{k=1}^K, i = 1, 2, \ldots, N$, learned by $\mathbf{H}(\cdot)$ should be near-binary codes. This implies that each element $\mathbf{B}_i^k$ of a feature vector should be close to either 1 or 0. Such an objective can be approximated by penalizing the learned feature to have elements close to 0.5

$$\ell_{pol}(\mathbf{B}_i) = \frac{1}{\frac{1}{2K}\sum_{k=1}^K \|\mathbf{B}_i^k - 0.5\|_2^2 + \epsilon}, \qquad (5)$$

in which ϵ is a small positive constant for avoiding divide-by-zero, and we use $\epsilon = 0.01$ in our experiments. We call such a loss in (5) as a *polarization* loss.

In addition, we expect that every bit in a binary code of length ι can contribute to the representation of individual

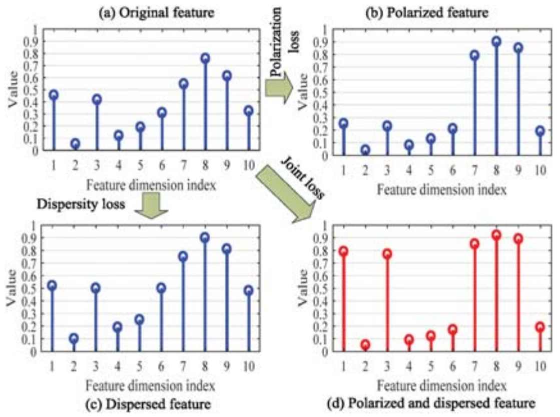


Fig. 4. The benefit of using the polarization loss and dispersity loss. (a) an original feature vector of real values in the range of $[0, 1]$, (b) the learned feature vector after using the polarization loss alone (i.e., each element is close to either 0 or 1), (c) the learned feature vector after using the dispersity loss alone (i.e., the elements are evenly distributed on the two sides of 0.5), and (d) the learned feature vector after jointly using the polarization and dispersity losses (i.e., near-binary elements are evenly distributed on the two sides of 0.5).

tattoos. In other words, each bit of the binary code is expected to have a 50 percent fire rate [42]. Such an objective can be approximated by constraining the average value of a learned feature to be 0.5, i.e.,

$$\ell_{dis}(\mathbf{B}_i) = \frac{1}{\iota}\sum_{k=1}^{\iota}\mathbf{B}_i^k - 0.5 \tag{6}$$

We call such a loss in (6) as a *dispersity* loss.

By minimizing the losses defined in (3), (5) and (6) jointly, the learned features are expected to be discriminative near-binary codes that are evenly distributed in the feature space (see Fig. 4). Such near-binary codes can be easily converted into a binary code utilizing a threshold function

$$\mathbf{B'}_i^k = \mathcal{T}(\mathbf{B}_i^k) = \begin{cases} 0 & \text{if } \mathbf{B}_i^k < 0.5 \\ 1 & \text{otherwise} \end{cases}. \tag{7}$$

Finally, we compute the distance between a query tattoo image and a gallery tattoo image using the Hamming distance of the binary vectors. In case more than one tattoos are detected from an image, i.e., u and v tattoos are detected from a query image and a gallery image, respectively, we compute $u \times v$ distances in total, and use the minimum distance as the final distance between the two tattoo images.

3.3 Network Structure

While our aim of joint tattoo detection and compact representation learning is well defined by (1), (3), (5), and (6), the feature sharing between detection and representation learning tasks is not simple.

We propose to perform joint tattoo detection and compact representation learning based on a Faster R-CNN model by embedding the above four losses into a single network. Specifically, a CNN such as AlexNet [30], VGG [69], or ResNet [70] can be used to extract the deep features that are to be shared by RPN and Fast R-CNN. The RPN and Fast R-CNN modules, and the tattoo regression loss defined in (1) in the proposed approach are the same as those in the original Faster R-CNN (see Fig. 3).

We establish joint compact representation learning by introducing a new compact representation learning (CRL) sub-network (see Fig. 3). CRL consists of an instance pooling layer (e.g., P_B), a sequence of fully connected (FC) layers (e.g., FC_{B_1} and FC_{B_2}), a latent layer with sigmoid activation (e.g., FC_l), and three sibling output layers (e.g., cls_loss, polarization, and dispersity). The instance pooling layer and the sequence of FC layers are the same as the RoI pooling and FC layers in Fast R-CNN, which compute a fixed-length feature vector from the shared feature map for each tattoo detection by Fast R-CNN, and optimize this feature w.r.t. the following CRL tasks. The latent layer (also an FC layer) with sigmoid activation is expected to generate near-binary representation $\mathbf{B}_i$ that will be binarized via (7) for large-scale search. The three sibling output layers model the corresponding constraints, i.e., polarization loss in (5), dispersivity loss in (6), and classification loss in (3), respectively (see Fig. 3). Overall, the three loss functions work in a multi-task way to perform CRL given a tattoo detection. We use hyper-parameters to control the balance between individual losses

$$\ell_J(\mathbf{B}_i) = \alpha\ell_{cls} + \beta\ell_{pol} + \gamma\ell_{dis}. \tag{8}$$

We set $\alpha = \beta = \gamma = 1$ based on empirical results. It should be noted that although there are K classes of tattoos in the training dataset, the number of outputs in both RPN and Fast R-CNN remains two (corresponding to background and tattoo). The number of outputs for classification loss layer in our CRL is K.

The proposed approach differs from [71], [72] in that: (i) [71] optimizes the feature representation for instance search by fine-tuning the detection network w.r.t. the query instances; however such features optimized for detection tasks may not be optimal for retrieval tasks. By contrast, the proposed approach jointly optimizes both detection and feature representation end-to-end; (ii) while the features used in [71], [72] are real-valued, the proposed approach learns compact binary features, which are scalable; (iii) the bounding boxes used for computing the features for instance matching in [72] are not accurate compared to the final bounding box estimates used by the proposed approach; however, as we will show in the experiments, accurate tattoo bounding boxes are important for improving the tattoo matching accuracy; (iv) while [72] studied person search with a gallery set containing about 6,978 images, the scalability of the proposed approach is studied with a gallery dataset containing more than 300K distracter tattoo images; and (v) while cross-modality instance search, i.e., sketch based tattoo search is studied in our work, the performance of [72] under a cross-modality matching scenario is not known.

3.4 Implementation Details

Data Augmentation. Compared with the large-scale databases for object detection [73] and image classification [74], the public-domain tattoo datasets such as Tatt-C [4], Flickr [10], and DeMSI [75] are of limited sizes (usually less than $10K$). The limited datase size poses additional challenges to the proposed approach, i.e., the risk of overfitting, particularly for our CRL module, which usually requires multiple tattoo

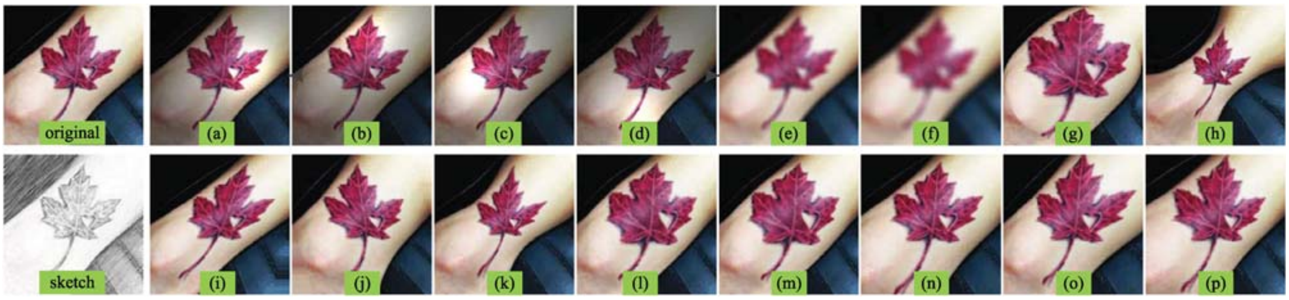


Fig. 5. An example of data augmentation for one tattoo image (referred to as "original" in the top row) in the training set to replicate various image acquisition conditions, e.g., (a-d) illumination variation, (e-f) image blur, (g-l) deformation, and (m-p) perspective distortion. A tattoo sketch is also generated for data augmentation (shown under the original tattoo).

images per class and a large number of tattoo classes to learn a robust model. Besides the commonly used data augmentation methods (e.g., random crop, translation, rotation, and reflection) [30], we have designed 16 additional transformations[8] to replicate the diversity of one tattoo instance caused by various acquisition conditions (see Fig. 5). In addition, we have also generated a tattoo sketch (see Fig. 5) for data augmentation so that the proposed approach can generalize to sketch-based tattoo retrieval task.

Network Training. We use ResNet-50 as our backbone network for shared feature learning, which is pretrained on ImageNet [74] for parameters initialization. We use a confidence threshold of 0.8 to filter the tattoo detections by Fast R-CNN, which corresponds to about one false detection for every ten images, on average. The filtered tattoo detections are fed into CRL. The reason why we use a relatively high threshold is to avoid feeding non-tattoo detections into CRL, wich may cause difficulty in network convergence. We use a learning rate of 10^{-4} during the fine-tuning of the pre-trained CRL. For the parameters of RPN and Fast R-CNN, we directly use the suggested values in [31], [66].

Since our approach performs tattoo detection and CRL jointly, the detection module usually takes one image as input (given a typical Titan X GPU), and outputs one detected tattoo, which is then used as the input to CRL. Thus, the batch size w.r.t. CRL is limited to one tattoo, which makes it difficult for CRL to converge. Such an issue cannot be resolved by just compiling a large training dataset or using data augmentation as in Fig. 5. To address this issue, we make use of preceding feature buffering [72] to assist in CRL training. In addition, we stitch multiple randomly selected training images into a single image (see Fig. 3), and use it as the input to our detection module so that it can output multiple detected tattoos. In this way, multiple detected tattoos can be used for training CRL, and thereby improving the training batch size. We have found such stitched tattoo images to be very for training our joint tattoo detection and CRL network. We also use online hard example mining (OHEM) [76] in Fast R-CNN to improve its robustness in detection blurred, partial, and tiny tattoos. All the tattoo images are scaled before they are input to the network so that the shorter edge between width and height is 600 pixels.

4 Experiments

4.1 Databases

There are only a limited number of tattoo databases in the public domain, such as Tatt-C [4], Flickr [10], and DeMSI [75]. These tattoo databases are used in the evaluations of our approach and comparisons with state-of-the-art.

Tatt-C. The NIST Tatt-C database was developed as an initial tattoo research corpus that addresses use cases representative of operational scenarios [4]. The tattoo versus non-tattoo classification dataset in Tatt-C contains 1,349 and 1,000 tattoo and non-tattoo images, respectively. The tattoo identification dataset in Tatt-C contains 157 and 215 probe and gallery images, respectively. A background dataset with 4,332 non-tattoo images was also used to populate the gallery set. The tattoo mixed-media dataset in Tatt-C, consisting of photos, sketches, and graphics, contains 181 and 272 probe and gallery images, respectively. A five-fold cross-validation was used for the identification experiment of each dataset in Tatt-C [4]. We also notice that the bounding-box annotations were provided for 7,526 tattoo images in Tatt-C, so we also report the tattoo detection accuracy using these image in Tatt-C. The tattoo images in the Tatt-C dataset contain variations of illumination, partial occlusion, and image blur (see Fig. 6a).

Flickr. The Flickr tattoo database contains 5,740 and 4,260 tattoo and non-tattoo images that were collected from Flickr [10]. We can notice that the ratio of the tattoo images to the non-tattoo images is similar to that of the Tatt-C database. While the tattoo images in the Tatt-C database were collected from an indoor environment, the images in the Flickr database were taken from both indoor and outdoor environment, with diverse viewpoints, poses, and complex backgrounds (see Fig. 6b). The Flickr database was original built for tattoo versus non-tattoo classification. We have extended the Flickr database by providing bounding-box annotations for each tattoo in the images.[9] Therefore, in our experiments, we are able to use the Flickr database to evaluate the tattoo detection (localization) performance.

DeMSI. The DeMSI dataset contains 890 tattoo images from the ImageNet [74] database. The boundary of each tattoo image was annotated for tackling the tattoo segmentation problem [75]. Since the tattoo images are from the ImageNet database, they are captured under an unconstrained scenario

[8] All the augmentations were performed leveraging the Photoshop plug-ins for MATLAB: https://helpx.adobe.com/photoshop/kb/downloadable-plugins-and-content.html

[9] We will put the bounding-box annotations for the Flickr database into the public domain.

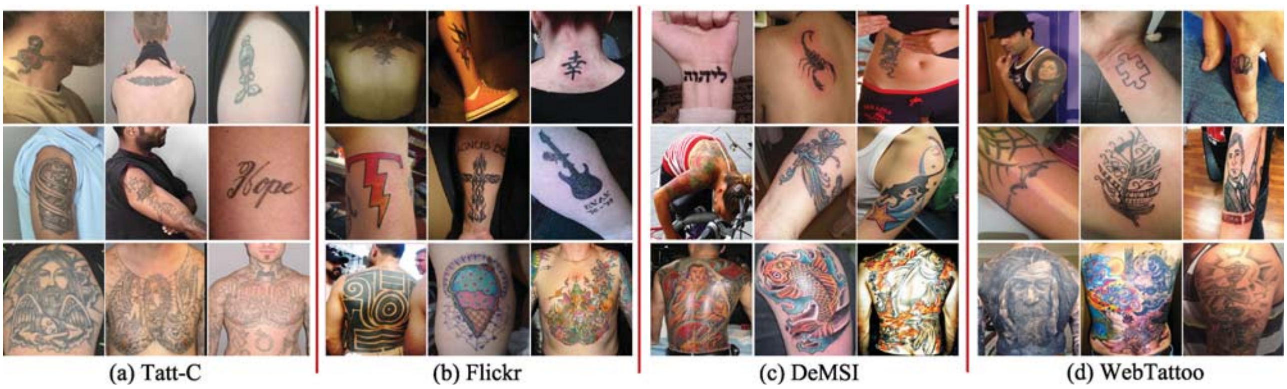
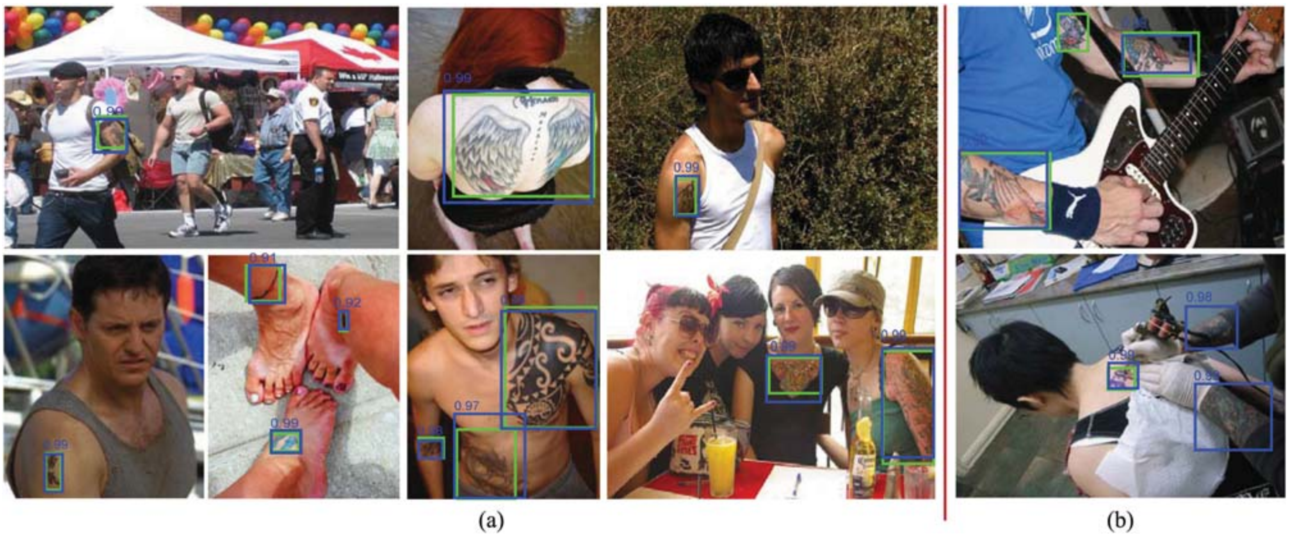

(a) Tatt-C (b) Flickr (c) DeMSI (d) WebTattoo

Fig. 6. Examples of tattoo images from the four tattoo image databases used in our experiments: (a) Tatt-C [5], (b) Flickr [10], (c) DeMSI [75], and (d) our WebTattoo dataset.



(a) (b)

Fig. 7. Examples of (a) good tattoo detections, and (b) poor tattoo detections by the proposed approach. The green and blue rectangles show the ground-truth tattoo bounding boxes and the detected tattoo bounding boxes, respectively. The numbers shown above the bounding boxes are the detection confidence scores.

(see Fig. 6c). The tattoo images from the DeMSI dataset are used together with our WebTattoo databases as the background images to populate the gallery dataset.

WebTattoo. We can notice that the above tattoo datasets are usually of limited sizes (less than $10K$). Although the NIST Tatt-E challenge is reported to have a much larger tattoo testing dataset collected from real application scenarios,[10] there is no evidence this dataset will be put into the public domain. To replicate the operational scenario of tattoo search at scale, we have compiled a large tattoo database (named as WebTattoo) by (i) combining the above three public-domain tattoo databases together, (ii) collecting over than 300K distracter tattoo images from the Internet (see Fig. 6d), and (iii) drawing 300 tattoo sketches by volunteers, who were asked to take a look at a tattoo image for one minute and then draw the tattoo sketch the next day (see an example of the tattoo sketch in Fig. 5).[11]

Based on the WebTattoo dataset, we use a semi-automatic approach to find the tattoo classes which have multiple tattoo images per class. Specifically, a ResNet-50 network pre-trained on ImageNet is first used for automatic tattoo feature extraction and clustering (we used k-means clustering [77]). The clusters are then manually verified to assure that each cluster contains only one class of tattoo images. Finally, we obtained about 600 tattoo classes, with nearly three tattoo images per class on average. We randomly choose about 1,400 tattoo images from 400 tattoo classes for training, and use tattoo images of the remaining 200 tattoo classes for testing. The training set is augmented using the method described in Section 3.4. We have manually annotated the tattoo bounding boxes for more than 78K tattoo images (original and augmented tattoo images) in total. Given such a large number of annotations, there are inevitably some missed tattoos by human workers (see Fig. 7b). However, such issues are not unique in our tattoo search task; they exist in many databases for individual computer vision tasks, such as face detection and recognition, person detection and recognition, etc. In addition to the data augmentation, we also use the 5,740 tattoo images from the Flickr dataset for training. Since no class label is provided for the tattoo images in Flickr, these tattoo images contribute only to the detection loss in the proposed

[10] https://www.nist.gov/programs-projects/tattoo-recognition-technology-evaluation-tatt-e

[11] We plan to put the WebTattoo dataset into the public-domain.

TABLE 2
Tattoo Detection (Localization) Performance of the Proposed Approach and the State-of-the-Art Methods on the WebTattoo Test and Tatt-C Datasets in Terms of Recall versus FPPI

Method	Recalls (in %) @ different FPPIs		
	0.01 FPPI Tatt-C/WebTatt	0.1 FPPI Tatt-C/WebTatt	1.0 FPPI Tatt-C/WebTatt
Sun et al. [22][1]	8/–	45/–	–/–
Faster R-CNN [32][2]	17.1/21.7	56.2/80.2	72.2/94.6
Proposed[2]	**45.9/27.5**	**61.7/87.1**	**80.0/95.5**

[1] *The results are from [21], in which a Tatt-C dataset with about 2,217 tattoo images was used.* [2]*Similar to [21], we trained a Faster R-CNN tattoo detector using the WebTattoo training set, and tested it on the Tatt-C (with 7,526 tattoo images) and WebTattoo test datasets. This is a cross-database testing scenario, which is more challenging than that used in [21]. We use an intersection-over-union (IoU) threshold of 0.5 between the detected and ground-truth bounding boxes.*

approach during network training. For each class of tattoos in the testing set, we randomly choose one tattoo image for the query, and use the remaining tattoo images for the gallery. Overall, we have 200 tattoo images in the query and 350 tattoo images in the gallery. About 300K WebTattoo images that are not present in the training, gallery, and query sets, are used as the distracter tattoo images to populate the gallery set, and replicate the large-scale tattoo search scenario. For the 300 pairs of tattoo sketches and the mated tattoo photos, we randomly choose 240 pairs for training, and the remaining 60 pairs for testing. For each pair of tattoo sketch and image, the tattoo sketch is used for query, and the tattoo image is used for gallery.

An operational tattoo dataset reported in [8] contains 327,049 tattoo images collected by the German police. Another operational tattoo dataset reported in [1] contains about 64,000 tattoo images, provided by the Michigan State Police. Our extended gallery set contains more than 300K tattoo images, which should reasonably replicate the operational tattoo search scenario.

4.2 Evaluation Metrics

The evaluations of the proposed approach and the comparisons with the state-of-the-art tattoo retrieval and identification methods cover the tasks of tattoo detection, identification, and large-scale search. For each task, we choose to use the widely used evaluation metric in the literature.

Tattoo Detection. We use the detection error trade-off (DET) curve to measure the tattoo detection performance, i.e., the recall versus false positives per image (FPPI). Given an intersection-over-union (IoU) threshold (we use 0.5) between the detected tattoo bounding boxes and the ground-truth tattoo bounding boxes, recall is defined as the fraction of detected bounding boxes with an IoU to the ground-truth larger than the threshold over the total amount of ground-truth bounding boxes.

Tattoo Search. We use the precision-recall curve to measure the tattoo search performance. Precision is the fraction of the mated tattoo images that have been retrieved over all the retrieved results for a given query tattoo. Recall, similar to that in the detection task, is the fraction of the mated tattoo images that have been retrieved over the total amount of mated tattoo images for a given query tattoo.

Tattoo Identification. We use the cumulative match characteristic (CMC) curve to measure the tattoo identification performance. Each point on CMC gives the fraction of the probe tattoo images that are correctly matched to their mated gallery images at a given rank.

4.3 Tattoo Detection

Since the proposed approach can perform tattoo detection and compact representation learning jointly, we first evaluate the tattoo detection performance of the proposed approach on the WebTattoo test and Tatt-C datasets. Specifically, we train our approach using the WebTattoo training set, and report the tattoo detection accuracy on the WebTattoo test and Tatt-C datasets. Since the Tatt-C dataset was primarily built for tattoo versus non-tattoo classification and tattoo identification tasks, only a limited number of published methods have reported tattoo detection performance on Tatt-C [21]. To provide more baseline performance, we train a Faster R-CNN tattoo detector used in [21] on the WebTattoo training dataset, and report its performance on the WebTattoo test and Tatt-C datasets. We should note that such a cross-database testing protocol is more challenging than the intra-database testing protocol used in [21].

Table 2 lists the tattoo detection performance of the proposed approach and the baseline methods. The state-of-the-art tattoo detection method in [21] reported about 45 percent recall @ 0.1FPPI on a Tatt-C dataset with about 2,000 tattoo images (one tattoo per image). The Faster R-CNN tattoo detector we trained gives 56.2 and 80.2 percent recalls at 0.1FPPI on the Tatt-C and WebTattoo test datasets, respectively. The performance of the Faster R-CNN tattoo detector we trained is much higher than that in [21], even though we are using a challenging cross-database testing protocol, and the Tatt-C subset we used for evaluation contains much more tattoo images than that was used in [21] (7,526 versus 2,000 tattoo images). The possible reason is that the WebTattoo training set contains more tattoo images than those used in the intra-database testing, which is helpful for training a deep learning based tattoo detector. In addition, our data augmentation can replicate the appearance variations existing in various tattoo images, and thus is helpful to improve the robustness of the tattoo detector in unseen scenarios. The proposed approach for joint tattoo detection and CRL achieves 61.7 and 87.1 percent recalls at 0.1FPPI on the Tatt-C and WebTattoo test datasets, respectively, which are much better than the state-of-the-art tattoo detectors. The results indicate that the proposed approach can leverage multi-task learning to achieve robust feature learning and detector modeling. Another baseline tattoo detection method in [20] used a Graphcut based method, and reported 70.5 percent precision @ 41.0 percent recall on a Tatt-C dataset with 6,308 tattoo images. Under the same evaluation metric, the proposed approach can achieve 99.0 percent precision @ 41.0 percent recall on the above Tatt-C dataset with 7,526 tattoo images.

Fig. 7 shows examples of good and poor tattoo detections by our approach on the WebTattoo database. We find that the proposed approach is quite robust to large pose and illumination variations as well as the diversity of tattoo

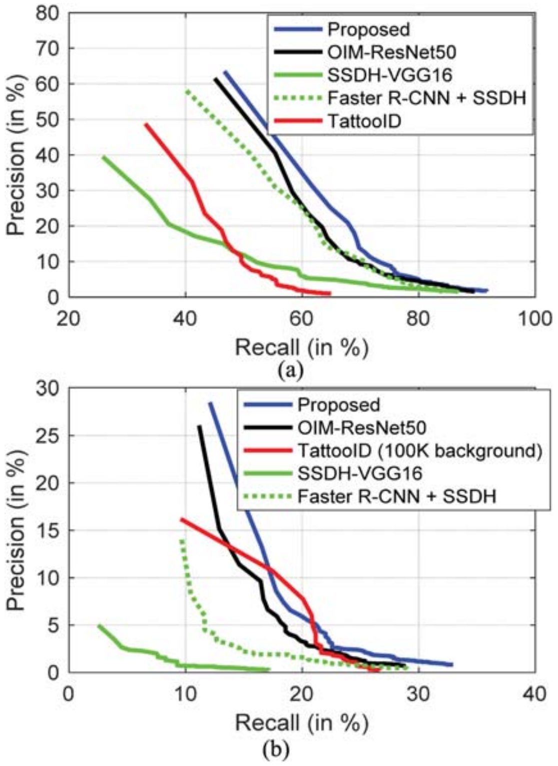


Fig. 8. Tattoo search performance (in terms of precision-recall) by the proposed approach and the state-of-the-art methods (TattooID [1], SSDH-VGG16 [42], Faster R-CNN + SSDH, and OIM-ResNet50 [73]) on the WebTattoo test dataset: (a) Without background tattoo images in the gallery set, and (b) with 300K background tattoo images in the gallery set; for TattooID, we report its tattoo search performance using an extended gallery set with only 100K background images because of its long running time.

categories. Some of the false detections by the proposed approach are due to the missed labeling of the tattoos (see the bottom tattoo image in Fig. 7b). However, we notice that detecting tiny tattoos that are easily confused with the background region remains a challenging problem.

4.4 Tattoo Search

Efficient tattoo search is important for scenarios, where the search must be operated in a large volume of raw images or video frames. We evaluate the our approach for tattoo search at scale, and provide comparisons with several state-of-the-art methods [1], [3], [42], [72]. For TattooID [1], [3], we reimplement it in Matlab because the early algorithm has been licensed to MorphoTrak[12]. For SSDH-VGG16 [42] and OIM-ResNet50 [72], we directly use the code provided with their papers, and train the models using the same training set as our approach. Since SSDH-VGG16 is not able to detect tattoos from an input image, we also consider another baseline, i.e., Faster R-CNN is applied for tattoo detection first, and then SSDH-VGG16 is used to extract compact features for the detected tattoos (Faster R-CNN + SSDH). We have tried several confidence thresholds (e.g., 0.3, 0.5, 0.7 and 0.9) for tattoo detection using Faster R-CNN, and finally chosen to use a threshold of 0.3, because of its good performance for the final tattoo search. For both SSDH-VGG16 and OIM-ResNet50, we use 256D feature representations as suggested in their papers. For our approach, we also use a 256-bit compact feature for fair comparisons. When the gallery size is too large, i.e., with 300K background tattoo images in the gallery, for efficiency, we compute the precision-recall using the top-100 retrieval results.

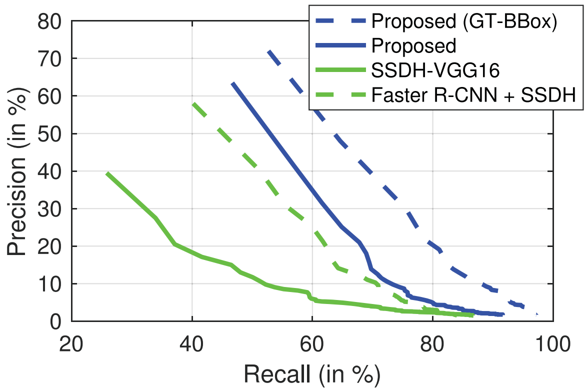


Fig. 9. The importance of using accurate tattoo bounding boxes for compact feature learning.

Fig. 8a shows the precision-recall curves of the proposed approach and the state-of-the-art methods for tattoo search without using the 300K background tattoo images to populate the gallery set. We are surprised to see that TattooID, a non-learning based matcher based on SIFT features, performs better than the deep learning based method SSDH-VGG16. The main reason is that SSDH-VGG16 alone is a holistic approach, which learns features from the entire tattoo images. Since many tattoo images contain large background regions around the tattoos, such a feature representation may capture more characteristics about the background regions than the tattoos, and thus leads to incorrect matches of tattoo images. This is also the reason why Faster R-CNN + SSDH achieves better performance than SSDH-VGG16. OIM-ResNet50, which is also a joint detection and feature learning method, is able to leverage the tattoo detection to reduce the influence of the background regions of the tattoos. As a result, the learned features could better represent the content of a tattoo, and achieves much higher tattoo search accuracy. However, OIM-ResNet50 extracts the features based on the region proposals of the tattoos instead of the final location estimations of the tattoos. Such a feature representation is less accurate than the proposed approach, which utilizes the final location estimations of the tattoos. The proposed joint tattoo detection and CRL approach performs better than all the baseline methods. The results suggest that multi-task learning used in our approach is helpful for learning more informative representation for tattoo search. In addition, the proposed approach leverages OHEM to improve the tattoo detection robustness, and stitched training images to increase the instance-level batch size during CRL. In Fig. 9, we also provide the performance of tattoo search using the ground-truth tattoo bounding boxes to extract our compact features. The results clearly show that using more accurate tattoo bounding boxes for compact feature learning does improve the tattoo matching accuracy. However, we also notice that the CRL module achieves only 72 percent rank-1 identification rate using the ground-truth tattoo bounding

[12] https://msutoday.msu.edu/news/2010/msu-licenses-tattoo-matching-technology-to-id-criminals-victims

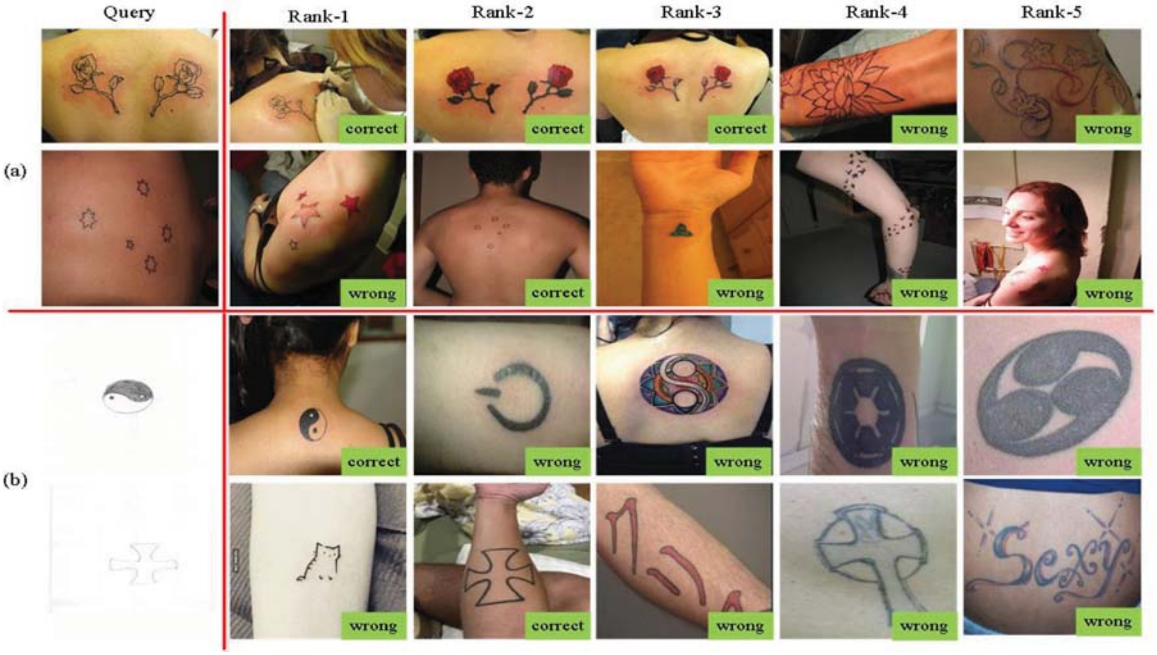


Fig. 10. Examples of tattoo image search results by the proposed approach using (a) tattoo photos and (b) tattoo sketches as queries. For each query tattoo image, the top-5 tattoo gallery images in the returned list are given.

boxes. The possible reason is that the small size of the training set (in terms of the number of tattoo classes and images) has limited the training of the CRL module. There is still room to improve the performance of the CRL module.

After we populate the gallery set using 300K background tattoo images, as expected, all the approaches report decreased tattoo search performance (Fig. 8b). Matching the query tattoo images with the complete 300K gallery using TattooID would take more than one week on CPU, so we use 100K background tattoo images for TattooID. Again, both OIM-ResNet50 and the proposed approach outperform TattooID and SSDH-VGG16 by a large margin, and our approach performs better than OIM-ResNet50. This suggests that the proposed approach remains effective under large-scale tattoo search scenarios.

We also evaluate the influence of different code lengths to the final tattoo search performance of our approach. As shown in Fig. 11, when we increase the code length to 512 bits, there is no performance improvement; instead, a minor precision drop is observed around 60 percent recall. If we reduce the code length to 128 bits, there will be a large performance drop of the precision. Therefore, we choose to use 256-bit compact features in all the experiments of our approach.

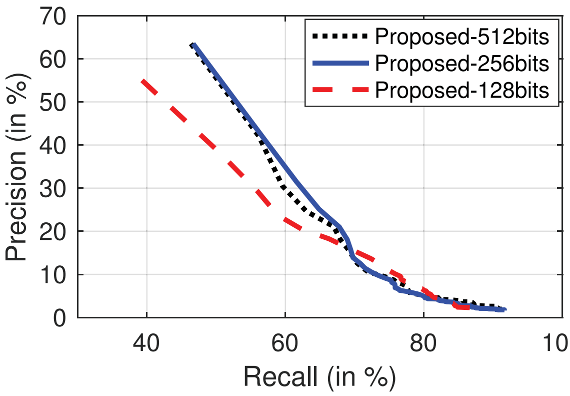


Fig. 11. The influence of the compact binary code length (in bit) to the tattoo search performance (in terms of precision-recall) by the proposed approach reported on the WebTattoo test dataset without using background tattoo images in the gallery set.

Fig. 10a shows examples of tattoo search results by our approach on the WebTattoo database, in which the top-5 matched gallery images are given for each query tattoo image. We can see that the proposed approach is robust against variations of body pose, illumination, and scale. Some of the incorrectly matched gallery tattoo images by the proposed approach show high visual similarity to the query tattoo image (see the second row in Fig. 10a).

4.5 Sketch Based Tattoo Search

In many scenarios, the surveillance image of the crime scene is not available, so the query is in the form of a sketch of a tattoo drawn based on the description provided by an eyewitness (see Fig. 10b). Therefore, it is important to evaluate the performance of a tattoo search system under a sketch based tattoo search scenario. SSDH-VGG16 [42], Faster R-CNN + SSDH, OIM-ResNet50 [72], and our method that are used in Section 4.4, are fine-tuned on the tattoo sketch training set consisting of 240 pairs of tattoo sketches and photos, and then evaluated on the tattoo sketch test set.

Fig. 12 shows the precision-recall curves of the proposed approach and the state-of-the-art methods for sketch-based tattoo search. As expected, as a cross-modality search problem, tattoo sketch-to-photo matching is much more challenging than tattoo photo-to-photo matching. Different from the observations in image-based tattoo search, SSDH-VGG16 performs better than TattooID in sketch-based tattoo search. The main reasons are two-fold: (i) TattooID detects much less SIFT keypoints from the tattoo sketches drawn on the papers than from the tattoo photos; (ii) the learning

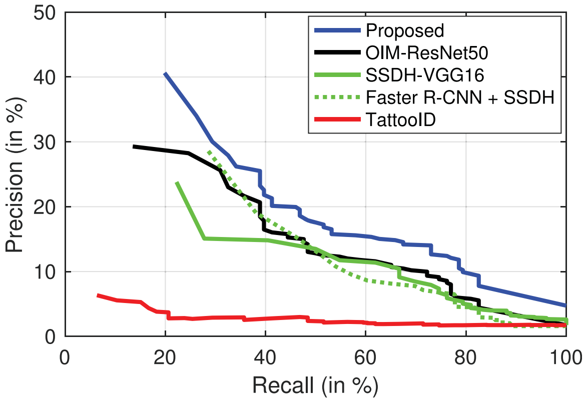


Fig. 12. Sketch based tattoo search performance (in terms of precision-recall) by the proposed approach and the state-of-the-art methods (TattooID [1], SSDH-VGG16 [43], Faster R-CNN + SSDH, and OIM-ResNet50 [73]) on the WebTattoo test dataset without background tattoo images in the gallery set.

based methods, such as SSDH-VGG16, are able to leverage the tattoo sketch-photo pairs to learn a feature representation that mitigates the modality gap between the tattoo sketches and photos. The methods that compute features from the detected tattoos (e.g., the proposed approach, OIM-ResNet50, and Faster R-CNN + SSDH) perform better than the methods that directly extract features from the holistic tattoo images. The proposed approach performs consistently better than the state-of-the-art methods in sketch based tattoo search. The results suggest that proposed joint tattoo detection and CRL approach has good generalization ability into the sketch-based tattoo search scenario.

Fig. 10b shows examples of sketch-based tattoo search results by our approach. Benefited from the joint tattoo detection in the proposed approach, our feature representation can reduce the influence of the background regions of the tattoos in the gallery set, and thus is able to match a tattoo sketch to its mated tattoo image at a low rank.

4.6 Tattoo Identification

Automatic tattoo identification techniques are usually utilized to generate a candidate suspect list, which is used for human or forensic analysis. While high rank-1 accuracy is ideal, success in these forensic recognition scenarios is generally measured by the accuracies from rank-1 to rank-100 [78]. Therefore, a number of the published tattoo identification methods reported their performance in terms of CMC curves covering rank-1 to rank-100 [1], [3], [7], [9], [19]. We report the CMC curves of the proposed approach and the state-of-the-art methods on the WebTattoo test dataset in Fig. 13. Again, the results show that the methods that compute features from the detected tattoos (e.g., the proposed approach, OIM-ResNet50, and Faster R-CNN + SSDH) perform better than the methods that directly extract features from the holistic tattoo images (e.g., SSDH and TattooID). However, the proposed approach, which leverages multi-task learning to perform joint tattoo detection and CFL, achieves the best accuracy. When the 300K background images are used to populate the gallery set, all the approaches are observed to have decreased identification accuracies, e.g., about 30 percent degradation at rank-1 identification accuracy. The proposed approach still performs better than the baselines in such a challenging scenario. These results indicate that the proposed approach has good generalization ability into the scenario of tattoo identification with a large gallery set.

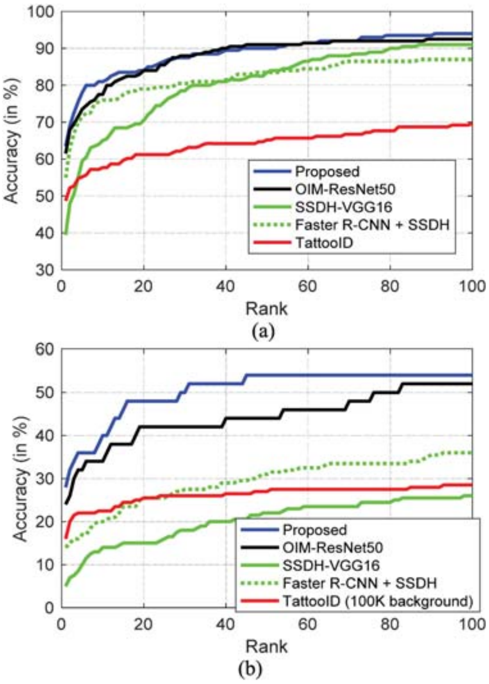


Fig. 13. Tattoo identification performance (in terms of CMC) by the proposed approach and the state-of-the-art methods (TattooID [1], SSDH-VGG16 [42], Faster R-CNN + SSDH, and OIM-ResNet50 [72]) on the WebTattoo test dataset: (a) Without 300K background tattoo images in the gallery set, and (b) with 300K background tattoo images in the gallery set; for TattooID, we report its tattoo search performance using an extended gallery set with only 100K background images because of its long running time.

On the public Tatt-C identification dataset, the proposed approach achieves 99.2 percent rank-1 identification accuracy. The MorphoTrak and the Purdue teams reported 99.4 and 98.7 percent rank-1 identification accuracies on the Tatt-C identification dataset. While the results by MorphoTrak are slightly better than ours, they used four folds of data of Tatt-C for training, and the fifth-fold data for testing. By contrast, our approach is trained on the WebTattoo training dataset, which is different from the tattoo images in Tatt-C.

In addition to the above evaluations, we also evaluate the generalization ability of the proposed approach in other instance-level retrieval tasks, such as on Paris [79] and Oxford [80], which contain challenging viewpoint and scale variations. Following the same testing protocol as the state-of-the-art method [42], our approach achieves 83.24 and 53.04 percent mAP on Paris and Oxford, respectively, which are comparable to the performance of state-of-the-art method [42] (83.87 and 63.79 percent mAP on Paris and Oxford, respectively). These results show that the proposed approach has good generalization ability to new application scenarios.

4.7 Ablation Study

We provide ablation studies of our approach in terms of three loss functions, i.e., (i) basic cross-entropy loss, (ii)

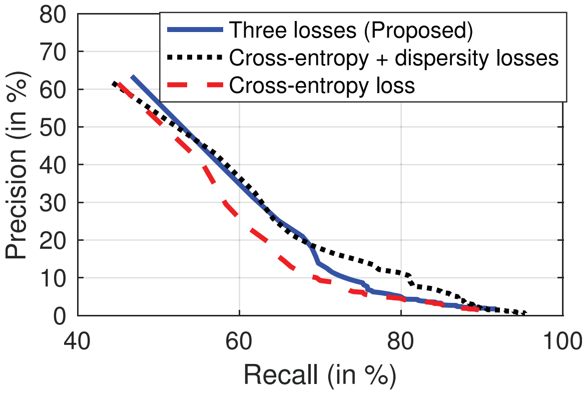


Fig. 14. Ablation studies involving three loss functions (cross-entropy, dispersity, and polarization) in CRL for tattoo image search on the WebTattoo test dataset without distracter images in the gallery set.

cross-entropy loss and dispersity loss, and (ii) all three losses together. The precision-recall curves calculated using the top-100 retrieval results of the three experiments are shown in Fig. 14. We can see that using dispersity loss together with the cross-entropy loss does not improve the rank-1 tattoo search accuracy compared to using cross-entropy loss alone, but it does improve the overall performance beyond rank-1. Jointly using all three losses leads to the best performance, particularly for the rank-1 tattoo search accuracy; this is important for practical applications. The reason why jointly using all three losses works better for compact feature learning is that while cross-entropy loss is helpful for generating real-valued codes that are discriminative between individual tattoo classes, dispersity and polarization losses assure the real-valued codes are near-binary and evenly distributed in the code space (see our explanations in Section 3.2 and Fig. 4).

4.8 Computational Cost

Our approach takes about 0.2 sec. in total to perform joint detection and CRL on a Titan X GPU. After obtaining the compact feature (256-bit), the average time of computing the Hamming distance of two 256-D binary codes is 0.06 ms on an Intel i7 3.6 GHz CPU without using bitwise operation based optimizations, which is 5 times faster than computing the cosine distance of two 256-D real-valued codes (0.3 ms on average). Such a difference in computational cost matters particularly for scenarios of tattoo search from huge volumes of surveillance video frames or handling multiple parallel searching requests. At the same time, comparisons with the state-of-the-art methods based on real-valued codes, e.g., [72], show that our compact binary codes of the same length can achieve better accuracy. Only a few of the published tattoo identification and retrieval methods have reported computational costs. For example, [15] reported an average of 24 sec. in comparing one query tattoo against 10K gallery tattoos after obtaining the SIFT features on an Intel Core2 2.66 GHz CPU, which is much slower than the proposed compact representation. We also profiled the tattoo detection time by a Faster R-CNN detector in [21] (using the same input image size as our approach) and the feature extraction time by a VGG-16 CNN network used in [42]. Tattoo detection and feature extraction per tattoo detection take 0.2 sec. and 0.04 sec., respectively. This indicates that the proposed approach is more efficient in real application scenarios, in which there are usually more than one tattoo detections per image.

5 Conclusions

This paper presents a joint detection and compact feature learning approach for tattoo image search at scale. While existing tattoo search methods mainly focus on matching cropped tattoos, the proposed approach models tattoo detection and compact representation learning in a single convolutional neural network via multi-task learning. The WebTattoo dataset consisting of 300K tattoo images was compiled from the public-domain tattoo datasets and images from the Internet. In addition, 300 tattoo sketches were created for sketch-based tattoo search to replicate the scenario where the surveillance image of the tattoo is not available. These datasets help evaluate the proposed approach for tattoo image search at scale and in operational scenarios. Our approach performs well on a number of tasks including tattoo detection, tattoo search at scale, and sketch-based tattoo search. The proposed data augmentation method is able to replicate various tattoo appearance variations, and thus is helpful to improve the robustness of the tattoo detector in unconstrained scenarios. Experimental results with cross-database testing protocols show that the proposed approach generalizes well to the unseen scenarios.

Acknowledgments

This research was supported in part by the Natural Science Foundation of China (grants 61732004, 61672496, and 61650202), Strategic Priority Research Program of CAS (grant XDB02070004), and External Cooperation Program of CAS (grant GJHZ1843). The preliminary work appeared in the Proceedings of the 6th International Conference on Biometrics (ICB), 2013 [9].

References

[1] J. Lee, R. Jin, A. K. Jain, and W. Tong, “Image retrieval in forensics: Tattoo image database application,” *IEEE MultiMedia*, vol. 19, no. 1, pp. 40–49, Jan. 2012.

[2] A. Bertillon, *Signaletic Instructions Including the Theory and Practice of Anthropometrical Identification*, Greenville, PA, USA: The Werner Company, 1896.

[3] A. K. Jain, J. Lee, and R. Jin, “Tattoo-ID: Automatic tattoo image retrieval for suspect and victim identification,” in *Proc. IEEE Pacific-Rim Conf. Multimedia*, 2007, pp. 256–265.

[4] M. L. Ngan, G. W. Quinn, and P. J. Grother, “Tattoo recognition technology - challenge (Tatt-C): Outcomes and recommendations,” NIST Interagency/Internal Report, Tech. Rep. 8078, pp. 1–29, 2015, https://ws680.nist.gov/publication/get_pdf.cfm?pub_id=919069

[5] R. McCabe, “Information technology: American national standard for information systems: Data format for the interchange of fingerprint, facial, & scar mark & tattoo (SMT) information,” NIST Special Publication, 500–245, pp. 1–80, 2000, https://docplayer.net/7370859-Nist-special-publication-500-245.html

[6] J. Lee, R. Jin, and A. K. Jain, “Rank-based distance metric learning: An application to image retrieval,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[7] S. T. Acton and A. Rossi, “Matching and retrieval of tattoo images: Active contour cbir and glocal image features,” in *Proc. IEEE Southwest Symp. Image Anal. Interpretation*, Mar. 2008, pp. 21–24.

[8] D. Manger, “Large-scale tattoo image retrieval,” in *Proc. 9th Conf. Comput. Robot Vis.*, May 2012, pp. 454–459.

[9] H. Han and A. K. Jain, "Tattoo based identification: Sketch to image matching," in *Proc. Int. Conf. Biometrics*, Jun. 2013, pp. 1–8.

[10] 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 *Proc. Int. Conf. Biometrics*, Jun. 2016, pp. 1–7.

[11] Y. Cao, C. Wang, L. Zhang, and L. Zhang, "Edgel index for large-scale sketch-based image search," in *Proc. IEEE Comput. Society Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 761–768.

[12] Y. Cao, H. Wang, C. Wang, Z. Li, L. Zhang, and L. Zhang, "MindFinder: Interactive sketch-based image search on millions of images," in *Proc. ACM Multimedia Int. Conf.*, Oct. 2010, pp. 1605–1608.

[13] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Nov. 2009.

[14] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain, "Matching composite sketches to face photos: A component-based approach," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 1, pp. 191–204, Jan. 2013.

[15] A. K. Jain, J. E. Lee, R. Jin, and N. Gregg, "Content-based image retrieval: An application to tattoo images," in *Proc. IEEE Int. Conf. Image Process.*, Nov. 2009, pp. 2745–2748.

[16] F. Li, W. Tong, R. Jin, A. K. Jain, and J. Lee, "An efficient key point quantization algorithm for large scale image retrieval," in *Proc. ACM Workshop Large-Scale Multimedia Retrieval Mining*, Oct. 2009, pp. 89–96.

[17] B. Heflin, W. Scheirer, and T. E. Boult, "Detecting and classifying scars, marks, and tattoos found in the wild," in *Proc. IEEE 5th Int. Conf. Biometrics: Theory Appl. Syst.*, Sep. 2012, pp. 31–38.

[18] M. J. Wilber, E. Rudd, B. Heflin, Y. Lui, and T. E. Boult, "Exemplar codes for facial attributes and tattoo recognition," in *Proc. Winter Conf. Appl. Comput. Vis.*, Mar. 2014, pp. 205–212.

[19] X. Xu and A. Kong, "A geometric-based tattoo retrieval system," in *Proc. 23rd Int. Conf. Pattern Recognit.*, Dec. 2016, pp. 3019–3024.

[20] J. Kim, H. Li, J. Yue, J. Ribera, E. J. Delp, and L. Huffman, "Automatic and manual tattoo localization," in *Proc. IEEE Symp. Technol. Homeland Security*, May 2016, pp. 1–6.

[21] Z. Sun, J. Baumes, P. Tunison, M. Turek, and A. Hoogs, "Tattoo detection and localization using region-based deep learning," in *Proc. 23rd Int. Conf. Pattern Recognit.*, Dec. 2016, pp. 3055–3060.

[22] X. Di and V. M. Patel, "Deep tattoo recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2016, pp. 119–126.

[23] X. Di and V. M. Patel, *Deep Learning for Tattoo Recognition*, B. Bhanu and A. Kumar, Eds, Berlin, Germany: Springer, 2017.

[24] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. II:1470–1477.

[25] L. Liu, J. Chen, P. Fieguth, G. Zhao, R. Chellappa, and M. Pietikäinen, "From BoW to CNN: Two decades of texture representation for texture classification," *Int. J. Comput. Vis.*, pp. 1–36, Nov. 2018.

[26] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[27] L. Liu and P. W. Fieguth, "Texture classification from random features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 574–586, Mar. 2012.

[28] Y. LeCun, Y. Bengio, and G. Hinton, "Learning representations by back-propagating errors," *Nature*, vol. 521, no. 14539, pp. 436–444, May. 2015.

[29] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1224–1244, May 2018.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[32] W. Zhou, H. Li, J. Sun, and Q. Tian, "Collaborative index embedding for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1154–1166, May 2018.

[33] L. Xie, Q. Tian, and B. Zhang, "Max-SIFT: Flipping invariant descriptors for web logo search," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 5716–5720.

[34] W. Zhou, Y. Lu, H. Li, and Q. Tian, "Scalar quantization for large scale image search," in *Proc. 20th ACM Int. Conf. Multimedia*, Oct. 2012, pp. 169–178.

[35] L. Xie, R. Hong, B. Zhang, and Q. Tian, "Image classification and retrieval are one," in *Proc. 5th ACM Int. Conf. Multimedia Retrieval*, 2015, pp. 3–10.

[36] L. Chi, H. Zhang, and M. Chen, "End-to-end spatial transform face detection and recognition," ArXiv e-prints, arXiv:1703.10818, pp. 1–9, Mar. 2017.

[37] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *Pro. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5525–5533.

[38] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," ArXiv e-prints, arXiv:1411.7923, pp. 1–9, Nov. 2014.

[39] M. Norouzi and D. J. Fleet, "Minimal loss hashing for compact binary codes," in *Proc. 28th Int. Conf. Int. Conf. Mach. Learn.*, Jun. 2011, pp. 353–360.

[40] V. E. Liong, J. Lu, G. Wang, P. Moulin, and J. Zhou, "Deep hashing for compact binary codes learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2475–2483.

[41] H. Jain, J. Zepeda, P. Perez, and R. Gribonval, "SUBIC: A supervised, structured binary code for image search," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 833–842.

[42] H. F. Yang, K. Lin, and C. S. Chen, "Supervised learning of semantics-preserving hash via deep convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 437–451, Feb. 2018.

[43] J. Lu, V. E. Liong, and J. Zhou, "Deep hashing for scalable image search," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2352–2367, May 2017.

[44] H. Jegou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 117–128, Jan. 2011.

[45] A. Babenko and V. Lempitsky, "Additive quantization for extreme vector compression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 931–938.

[46] T. Zhang, C. Du, and J. Wang, "Composite quantization for approximate nearest neighbor search," in *Proc. 31st Int. Conf. Int. Conf. Mach. Learn.*, Jun. 2014, pp. II:838–846.

[47] X. Wang, T. Zhang, G. Qi, J. Tang, and J. Wang, "Supervised quantization for similarity search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2018–2026.

[48] T. Yu, Z. Wang, and J. Yuan, "Compressive quantization for fast object instance search in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 726–735.

[49] P. Indyk and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in *Proc. 30th Annu. ACM Symp. Theory Comput.*, May 1998, pp. 604–613.

[50] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on P-stable distributions," in *Proc. 20th Annu. Symp. Comput. Geometry*, Jun. 2004, pp. 253–262.

[51] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. 21st Int. Conf. Neural Inf. Process. Syst.*, Dec. 2009, pp. 1753–1760.

[52] B. Kulis and K. Grauman, "Kernelized locality-sensitive hashing for scalable image search," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2130–2137.

[53] W. Liu, J. Wang, S. Kumar, and S.-F. Chang, "Hashing with graphs," in *Proc. 28th Int. Conf. Int. Conf. Mach. Learn.*, Jun. 2011, pp. 1–8.

[54] W. Kong and W. Li, "Isotropic hashing," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, Dec. 2012, pp. 1646–1654.

[55] Q. Jiang and W. Li, "Scalable graph hashing with feature transformation," in *Proc. 24th Int. Conf. Artif. Intell.*, Jul. 2015, pp. 2248–2254.

[56] B. Kulis, P. Jain, and K. Grauman, "Fast similarity search for learned metrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2143–2157, Dec. 2009.

[57] W. Liu, J. Wang, R. Ji, Y. G. Jiang, and S. F. Chang, "Supervised hashing with kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2074–2081.

[58] Y. Lin, R. Jin, D. Cai, S. Yan, and X. Li, "Compressed hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 446–451.

[59] F. Shen, C. Shen, W. Liu, and H. T. Shen, "Supervised discrete hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 37–45.

[60] J. Gui, T. Liu, Z. Sun, D. Tao, and T. Tan, "Fast supervised discrete hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 490–496, Feb. 2018.

[61] R. Salakhutdinov and G. Hinton, "Learning a nonlinear embedding by preserving class neighbourhood structure," in *Proc. 11th Int. Conf. Artif. Intell. Statistics*, 2007, pp. 412–419.

[62] B. Kulis and T. Darrell, "Learning to hash with binary reconstructive embeddings," in *Proc. 22nd Int. Conf. Neural Inf. Process. Syst.*, Dec. 2009, pp. 1042–1050.

[63] W.-J. Li, S. Wang, and W.-C. Kang, "Feature learning based deep supervised hashing with pairwise labels," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, Jul. 2016, pp. 1711–1717.

[64] J. Wang, W. Liu, S. Kumar, and S. F. Chang, "Learning to hash for indexing big data - a survey," *Proc. IEEE*, vol. 104, no. 1, pp. 34–57, Jan. 2016.

[65] L. Chi and X. Zhu, "Hashing techniques: A survey and taxonomy," *ACM Comput. Surv.*, vol. 50, no. 1, pp. 11:1–36, Apr. 2017.

[66] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[67] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[68] P. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Ann. Operations Res.*, vol. 134, no. 1, pp. 19–67, Feb. 2005.

[69] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, May 2015, pp. 1–14.

[70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[71] A. Salvador, X. Giró-i-Nieto, F. Marqués, and S. Satoh, "Faster R-CNN features for instance search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2016, pp. 394–401.

[72] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3376–3385.

[73] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vision*, Sept. 2014, pp. 740–755

[74] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F.-F. Li, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[75] T. Hrkać, K. Brkić, and Z. Kalafatić, "Tattoo detection for soft biometric de-identification based on convolutional neural networks," in *Proc. OAGM-ARW Joint Workshop*, May 2016, pp. 131–138.

[76] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 761–769.

[77] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999.

[78] A. K. Jain, B. Klare, and U. Park, "Face matching and retrieval in forensics applications," *IEEE MultiMedia*, vol. 19, no. 1, pp. 20–28, Jan. 2012.

[79] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[80] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

Hu Han received the BS degree in computer science from Shandong University, in 2005, and the PhD degree in computer science from ICT, CAS, in 2011, respectively. He is an associate professor of the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS). Before joining the faculty at ICT, CAS in 2015, he has been a Research Associate at PRIP lab in the Department of Computer Science and Engineering, Michigan State University, and a visiting researcher at Google in Mountain View. His research interests include computer vision and pattern recognition with focus on bio-perception oriented intelligent computing. He is a member of the IEEE.

Jie Li received the BS degree from Qingdao University, in 2017, and he is working toward the MS degree in the School of Artificial Intelligence, University of Chinese Academy of Sciences. He is now a visiting student at ICT, CAS. His research interests include computer vision and pattern recognition with focus on bio-perception oriented intelligent computing.

Anil K. Jain is a University distinguished professor with the Department of Computer Science and Engineering, Michigan State University. His research interests include pattern recognition and biometric authentication. He served as the editor-in-chief of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* (1991-1994). He served as a member of the United States Defense Science Board and The National Academies committees on Whither Biometrics and Improvised Explosive Devices. He has received Fulbright, Guggenheim, Alexander von Humboldt, and IAPR King Sun Fu awards. He is a member of the National Academy of Engineering and foreign fellow of the Indian National Academy of Engineering. He is a fellow of the AAAS, ACM, IAPR, SPIE, and IEEE.

Shiguang Shan is a professor of ICT, CAS, and the deputy director with the Key Laboratory of Intelligent Information Processing, CAS. His research interests cover computer vision, pattern recognition, and machine learning. He has authored more than 200 papers in refereed journals and proceedings in the areas of computer vision and pattern recognition. He was a recipient of the China's State Natural Science Award in 2015, and the China's State S&T Progress Award in 2005 for his research work. He has served as the Area Chair for many international conferences, including ICCV'11, ICPR'12, ACCV'12, FG'13, ICPR'14, ACCV'16, FG'18, ACCV'18, BTAS'18, and CVPR'19. He is an associate editor of several journals, including the *IEEE Transactions on Image Processing*, the Computer Vision and Image Understanding, the Neurocomputing, and the Pattern Recognition Letters. He is a senior member of the IEEE.

Xilin Chen is a professor of ICT, CAS. He has authored one book and more than 200 papers in refereed journals and proceedings in the areas of computer vision, pattern recognition, image processing, and multimodal interfaces. He was a recipient of several awards, including the China's State Natural Science Award in 2015, the China's State S&T Progress Award in 2000, 2003, 2005, and 2012 for his research work. He is currently an associate editor of the *IEEE Transactions on Multimedia, and Journal of Visual Communication and Image Representation*, a leading editor of the Journal of Computer Science and Technology, and an associate editor-in-chief of the Chinese Journal of Computers, and Chinese Journal of Pattern Recognition and Artificial Intelligence. He served as an Organizing Committee member for many conferences, including general co-chair of FG13/FG18, Local Chair of ICME07, ACM MM09, and ICIP17, and Finance Chair of ISCAS13. He is/was an Area Chair of CVPR 2017/2019, and ICCV 2019. He is a fellow of the IEEE, IAPR, and CCF.