Received: 19 July 2021 | Revised: 28 February 2022 | Accepted: 10 March 2022

DOI: 10.1049/bme2.12071

IET Biometrics

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Using double attention for text tattoo localisation

Xingpeng Xu[1] | Shitala Prasad[2] | Kuanhong Cheng[3] | Adams Wai Kin Kong[1]

[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

[2]Visual Intelligence Department/Advanced Manufacturing and Engineering Division, Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore, Singapore

[3]School of Physics and Optoelectronic Engineering, Xidian University, Xi'an, China

Correspondence

Xingpeng Xu, School of Computer Science and Engineering, Nanyang Technological University, 639798 Singapore, Singapore.
Email: xpxu@ntu.edu.sg

Funding information

Ministry of Education, Singapore, Grant/Award Number: Academic Research Fund Tier 1, RG21/19-(S)

Abstract

Text tattoos contain rich information about an individual for forensic investigation. To extract this information, text tattoo localisation is the first and essential step. Previous tattoo studies applied existing object detectors to detect general tattoos, but none of them considered text tattoo localisation and they neglect the prior knowledge that text tattoos are usually inside or nearby larger tattoos and appear only on human skin. To use this prior knowledge, a prior knowledge-based attention mechanism (PKAM) and a network named Text Tattoo Localisation Network based on Double Attention (TTLN-DA) are proposed. In addition to TTLN-DA, two variants of TTLN-DA are designed to study the effectiveness of different prior knowledge. For this study, NTU Tattoo V2, the largest tattoo dataset and NTU Text Tattoo V1, the largest text tattoo dataset are established. To examine the importance of the prior knowledge and the effectiveness of the proposed attention mechanism and the networks, TTLN-DA and its variants are compared with state-of-the-art object detectors and text detectors. The experimental results indicate that the prior knowledge is vital for text tattoo localisation; The PKAM contributes significantly to the performance and TTLN-DA outperforms the state-of-the-art object detectors and scene text detectors.

KEYWORDS

attention mechanism, forensics, tattoo localisation, text tattoo localisation, visual identification

1 | INTRODUCTION

Tattoo is a very informative soft biometric, which carries rich personal information in many cases. Tattoos are usually not too small and often have describable contents such as text, symbols and some meaningful patterns and objects. Hence, they can be easily observed, remembered and described by witnesses. Figure 1 shows some photos of persons with tattoos. Because of these characteristics, tattoos have been widely used by law enforcement agencies for forensic investigation, such as criminal and victim identification. According to the statistic in Ref. [1], 85% of prisoners under 35 have tattoos and 75% of the inmates who are re-incarcerated have tattoos. Another statistic [2] reports that 36% of Americans between the age of 18 and 29 have at least one tattoo. Obviously, tattoos play an important role in criminal and victim identification.

Text tattoo, as a subcategory of tattoos, provides additional cues such as names, important dates and other personal information as shown in Figure 2. In most cases, each character in a text tattoo has a similar style, colour and font. They form a coherent pattern, which is different from general tattoos, where different parts can have different styles and colours. The bounding boxes of text tattoos are more likely rectangular but the shape of bounding boxes of general tattoos are very diverse. These cues can be important and useful during the forensic investigations. Therefore, a method that is able to localise text tattoos is essential. In addition, localisation is an essential step for automatically recognising text in the tattoos, which allows using text-based searching on text tattoos. In this study, we concentrate on text tattoo localisation. There are some important characteristics among general objects, tattoos and text tattoos, which can be exploited for text tattoo localisation. Tattoos, including text tattoos can only appear on the skin but not on other objects and text tattoos usually appear as a small part of a bigger tattoo with similar colour as shown in Figure 2. If objects and humans can be detected or highlighted, tattoo localisation will become more effective and similarly, if tattoo can be detected or highlighted, text tattoo localisation

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *IET Biometrics* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

20474946, 2022, 3, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/bme2.12071, Wiley Online Library on [16/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

FIGURE 1 Sample images in the NTU Tattoo V2 dataset

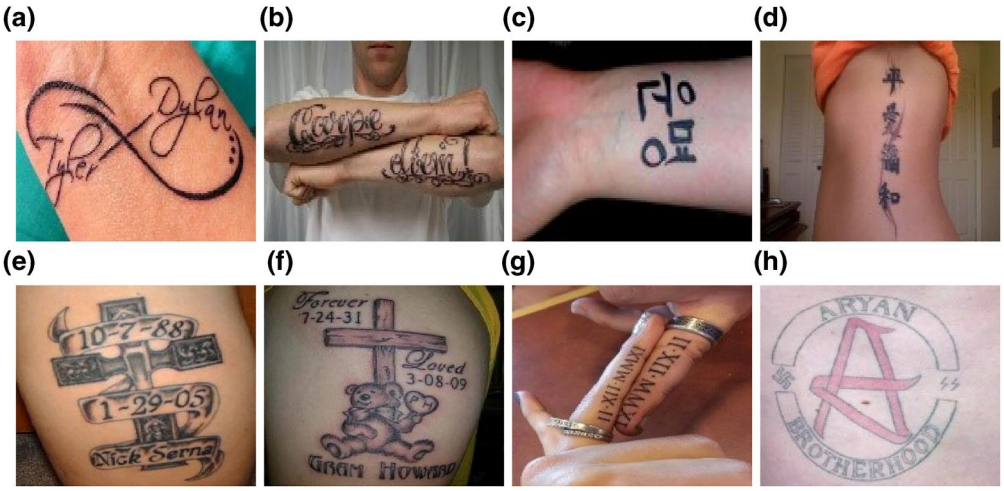

FIGURE 2 Sample text tattoos show important personal information: (a, b, e and f) names, (e, f and g) dates and (h) gang

will also become more effective. To make use of this prior knowledge, a prior knowledge-based attention mechanism (PKAM), which can be easily plugged into the backbones of standard detectors, is proposed. To use the relationship among objects, humans, general tattoos and text tattoos, the PKAM is applied twice in the proposed Text Tattoo Localisation Network Based on Double Attention (TTLN-DA). In addition to TTLN-DA, its two different variants are also developed for comparison and for fulfiling different demands.

Text tattoo can also be considered as a part of scene text, but the features of text tattoo and scene text are different in some ways. First, scene text is likely generated by computer using a set of popular fonts. Text tattoos, on the other hand, are designed and 'written' by the tattooist and have diverse styles. Second, scene text generally appears on a flat surface of an object, while text tattoos are on elastic surface, that is, human skin. Third, scene text normally has regular shape and each character has similar size in each word, while the shape of text tattoo can be arbitrary and each character in the text tattoo can have different sizes. Besides those differences, text tattoos can be inside another tattoo pattern or surrounded by other tattoos. Figure 3 shows some examples.

The existing public databases for tattoo localisation are small and none of them is for text tattoo localisation study. Thus, a new tattoo dataset named NTU Tattoo V2 and a text tattoo dataset named NTU Text Tattoo V1 containing, respectively, 87,282 general tattoo images and 9067 text tattoo images are established. NTU Tattoo V2 is the largest dataset for tattoo localisation study and NTU Text Tattoo V1 is the largest dataset for text tattoo localisation study.

20474946, 2022, 3, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/bme2.12071, Wiley Online Library on [16/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

FIGURE 3 The first row: sample images of text tattoos. The second row: sample images of scene text

The main contributions can be summarised in four fold:

- A new tattoo dataset and a new text tattoo dataset are established. To the best of our knowledge, these datasets are the largest tattoo and text tattoo datasets, which consist of much more challenging images compared with other existing public datasets.
- A novel method called TTLN-DA, which is especially designed for exploiting the relationships among objects, tattoos and text tattoos, is proposed. Two variants are also designed for different demands.
- A prior knowledge-based attention mechanism (PKAM), which extensively uses the relationship among objects, human, tattoo and text tattoo, is proposed.
- The experimental results demonstrate the importance of the relationships among objects, tattoos and text tattoos, which were totally neglected, and TTLN-DA outperforms the state-of-the-art object detectors and text detectors.

The remainder of this paper is structured as follows. Section 2 presents a literature review on object, tattoo and text localisation. Section 3 introduces the NTU Tattoo V2 and NTU Text Tattoo V1 datasets. Section 4 describes the proposed TTLN-DA, including PKAM and its variants. Section 5 gives the implementation details. Section 6 reports the experimental results and comparisons and Section 7 offers some conclusive remarks.

2 | RELATED WORK

Text tattoo localisation can be considered as a particular case of tattoo localisation, scene text localisation or object localisation, which share some similarities but have some dissimilarities. This section reviews representative methods from these three groups of techniques.

2.1 | Tattoo localisation

Tattoo localisation is defined as localising tattoo regions in images. In computer vision, researchers use the term detection for classifying and localising objects in images. However, in some tattoo recognition work [3], tattoo detection is defined as determining an image contains a tattoo or not. To avoid ambiguity, we use the term localisation in our work. For advancing research and development into tattoo localisation technology, the U.S. National Institute of Standards and Technology (NIST) held a challenge (Tatt-C) in 2015 [4] and released a tattoo dataset. Another competition (Tatt-E) [5] was held by NIST in 2018, and the competition results were published as an internal report. Shan et al. [6] hosted a robust tattoo localisation and retrieval competition (RTDRC) in 2019. Tatt-C, Tatt-E and RTDRC pinpointed the importance of tattoo localisation for forensic applications.

Some patch-based approaches were developed to localise tattoos from digital images. Huynh et al. [7] used a decision forest taking patch features to detect tattoos from full body images captured by a system designed for prisoners in a controlled environment. Hrkać et al. [8] used a deep convolutional neural network to discriminate the tattoo patches and non-tattoo patches and grouped tattoo patches into blobs. Those tattoo patches were cropped from close-up tattoo images. Sun et al. [9] proposed TATT-RBDL, which uses Faster R-CNN [10] to perform tattoo localisation when bounding box information is provided for training.

In forensic investigation, the collected images are more likely to be taken from uncontrolled environments with diverse poses, viewpoints and complex backgrounds. Some tattoo localisation studies were proposed to handle images collected in these environments. Heflin et al. [11] introduced a methodology that uses a graph-based visual saliency model (GBVS) [12] and GrabCut [13] for localising tattoos found in unconstrained images. Han et al. [14] proposed an efficient tattoo localisation method that is able to learn tattoo localisation and representation in a single CNN via multi-task learning. Chowdhury et al. [15] combines a skin segmentation procedure with a deformable convolution and inception (DCINN)-based scene text detector. However, the hard-coded thresholds in their segmentation procedure make it hard to apply to images with diverse illumination conditions. Their detection network was based on scene text detector. However, our experimental results (Table 3) show that the features of text tattoos are different from the scene text. Therefore, a scene text detector architecture may not be a good choice for text tattoo localisation task.

2.2 | Text localisation

Text localisation in natural scene images has been studied for decades and many image processing-based [16–20] and deep learning-based [21–24] methods have been proposed.

20474946, 2022, 3, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/bme2.12071, Wiley Online Library on [16/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Some text localisation networks are worth paying attention to. Zhou et al. [25] proposed a scene text detector, called EAST, which predicts word and line-level text instances from natural scene images by using a single neural network. Shi et al. [26] proposed an oriented text localisation method, named SegLink. SegLink decomposes text into two elements: segments and links. Segment is one of the oriented bounding box of a text line, while the link connects two adjacent segments that belong to the same word or text line. Deng et al. [27] proposed an instance segmentation-based method called PixelLink, which segments text pixels first and then generates bounding box from the segmentation results.

EAST, SegLink and PixelLink detect text with regular shape from natural scene images. To detect text with arbitrary shape, Wang et al. [28] proposed an efficient and accurate arbitrary-shaped text detector, called Pixel Aggregation Network (PAN). PAN uses a lightweight backbone network to extract the features of inputted images and enhance the extracted feature via Feature Pyramid Enhancement Module (FPEM) and Feature Fusion Module (FFM). Xie et al. [29] proposed a supervised pyramid context network (SPCN) to detect scene text. SPCN uses the feature pyramid network (FPN) [30] and text context modules to extract features and then utilises a Mask R-CNN [31] to predict the bounding box of scene text. To avoid the false localisation when two text instances are close to each other, Wang et al. [32] proposed a Progressive Scale Expansion Network (PSENet), which can detect multiple text instances with arbitrary shapes. PSENet uses a FPN to extract features at different scales and applies an image processing-based approach, named scale expansion algorithm, to calculate the region of text instances. Liao et al. [33] proposed a differentiable binarisation module to improve the performance of text localisation. Ye et al. [34] proposed the TextFuseNet to exploit the use of richer features fused for text localisation. TextFuseNet perceives text from character, word and global level of feature representation and uses a novel text representation fusion technique to achieve robust arbitrary text localisation. Liu et al. [35] used a parameterised Bezier curve to adaptively fit the arbitrarily-shaped text and proposed the ABCNet, which uses the Bezier Align layer for extracting accurate convolution features of a text instance with arbitrary shapes and significantly improving the precision.

2.3 | Object localisation

Text tattoo localisation can be considered as a sub-task of general object localisation. The backbones of some tattoo localisation and text localisation studies are deep object detectors, such as Faster R-CNN [10] and Mask R-CNN [31]. Therefore, some latest object detectors are worth paying attention to. TridentNet, the most recent state-of-the-art object detector, is proposed by Li et al. [36]. TridentNet is able to generate scale-specific feature maps with a uniform representational power by using a multi-branch architecture. Each branch shares the same transformation parameters but with different size of receptive fields. A scale-aware training scheme is adopted to specialise each branch by sampling object instances of proper scale for training. Wang et al. [37] proposed Cross Stage Partial Network (CSPNet) to mitigate the problem that other studies require heavy inference computations from the network architecture perspective. CSPNet reduces computations by 20% with equivalent or even superior accuracy. Qiao et al. [38] proposed Detectors that improves the performance of object localisation by using recursive feature pyramid at the macro level and switchable atrous convolution at the micro level.

3 | DATASETS

A text tattoo dataset is essential for training the proposed text tattoo localisation networks. To the best knowledge of the authors, there is no text tattoo dataset released for public access yet. For this work, we established a text tattoo dataset named NTU Text Tattoo V1, which contains 9067 images and 10,027 text tattoo instances.

The proposed text tattoo localisation networks utilise the relationship between general tattoo and text tattoo and therefore, a large general tattoo dataset is also required. Existing open access tattoo datasets (Tatt-C [4], DeMSI [39], NTU Tattoo V1 [40] and WebTattoo [14]) contain very limited number of tattoo images for tattoo localisation task. Table 1 summarises the size of each public tattoo datasets. For developing stronger localisation algorithms and accurately evaluating them, a new dataset, NTU Tattoo V2, is established, which consists of 78,215 non-text tattoo images collected from the Internet and 9067 text tattoo images from NTU Text Tattoo V1. The images taken in various environments, including indoor and outdoor, have diverse quality, sizes and illumination conditions. The entire NTU Tattoo V2 dataset has 92,838 manually annotated bounding boxes of non-text tattoos and 10,027 manually annotated bounding boxes of text tattoos. Figure 1 shows some sample images in the NTU Tattoo V2 dataset. The full dataset will be available for public access[1] once the paper is published.

4 | TEXT TATTOO LOCALISATION NETWORK

This section elaborates the proposed Text Tattoo Localisation Network based on Double Attention (TTLN-DA) and its variants. An attention mechanism based on the prior

TABLE 1 Comparisons of different tattoo datasets

Dataset	Image no.	Remarks
Tatt-C	7526	Indoor environments
DeMSI	890	Unconstrained environments
NTU tattoo V1	5740	No bounding box annotations
WebTattoo	5000	Indoor and unconstrained environments

20474946, 2022, 3, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/bme2.12071, Wiley Online Library on [16/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

knowledge of text tattoo is first described. Then, TTLN-DA and its variants based on the attention model are given.

4.1 | Prior knowledge-based attention mechanism

Inspired by the human visual behaviour, hierarchical attention mechanisms have been proposed to improve object localisation performance. Attention mechanisms are used to enhance important features and suppress unnecessary ones. Most popular mechanisms can be divided into three types: spatial mechanism [41–46], channel mechanism [47] and hybrid mechanism [48]. These three types of attention mechanisms enhance the features based on the object saliency and the importantness of the features. They neglect the relationship between objects from prior knowledge. To use the relationships among general objects, humans, tattoos, and text tattoos to improve text tattoo and general tattoo localisation performance, a prior knowledge-based attention mechanism (PKAM) is proposed.

The PKAM is designed to enhance the features within the portion of a particular object in the images. In this study, the PKAM is used to highlight tattoos such that text tattoos can be detected effectively. It can also be used to highlight objects and humans such that tattoos can be detected effectively. The PKAM takes two inputs: one is the original image I_t, containing objects, humans, tattoos and text tattoos and the other is the features f_t^R extracted by a detector. For highlighting tattoos, the detector is trained on a tattoo dataset to extract tattoo features, but for highlighting objects and humans, the detector is trained on a general image database to extract corresponding features. More clearly, f_t^R in Figures 4 and 5 can represent tattoo features and object and human features, depending on the detector. The PKAM has two versions named spatial PKAM and channel PKAM. The spatial PKAM is based on element-wise operators to fuse input images with the features from the detector and the channel PKAM is based on channel-wise operations to fuse input images with the features from the detector.

Spatial PKAM. Let the dimensions of f_t^R be $m \times n \times c$ and the dimension of I_t be $h \times w \times 3$. The spatial PKAM

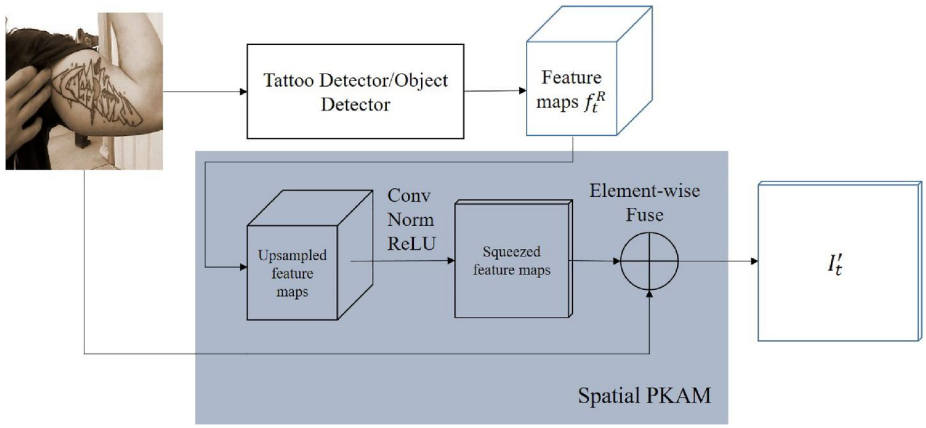

FIGURE 4 The prior knowledge-based attention mechanism (PKAM). I_t is a text tattoo image. f_t^R is the tattoo features or object features extracted by the tattoo detector or object detector, respectively. I_t' is the text tattoo image enhanced by PKAM. The highlighted block illustrates the architecture of spatial PKAM

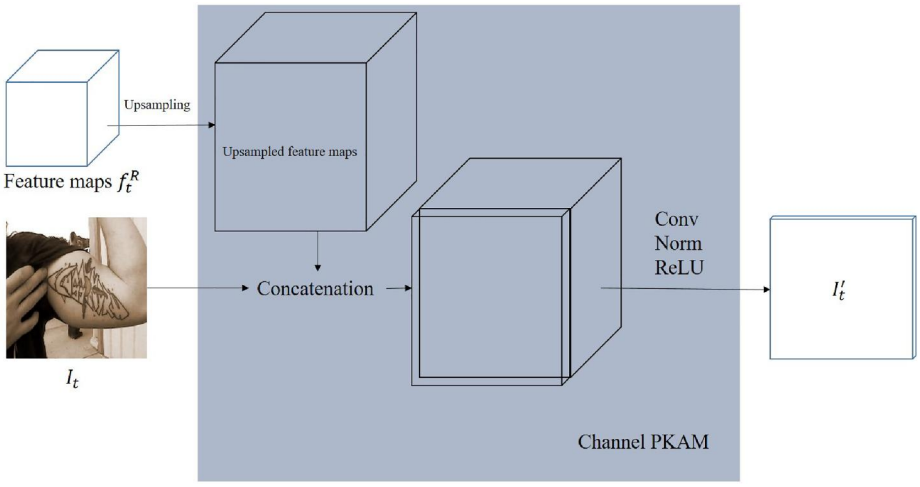

FIGURE 5 The architecture of channel prior knowledge-based attention mechanism

20474946, 2022, 3, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/bme2.12071, Wiley Online Library on [16/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

takes the features f_t^R and upsamples it to $h \times w \times c$ through the bilinear upsampling algorithm. Thus, its spatial resolution is same as the input image I_t. For the spatial PKAM (see Figure 4), a series of convolutional operations are applied to reduce the number of the channels of upsampled feature from c to 3. After the convolution operations, normalisation and ReLU are applied and the squeezed features and I_t are fused through element-wise sum operation to obtain I'_t, which is submitted to the next CNN block for extracting tattoo or text tattoo features.

Channel PKAM. The channel PKAM allows the network to learn how to select the important channels for each feature map locations. Figure 5 illustrates the architecture of the channel PKAM. Inside the channel PKAM, f_t^R is first upsampled, same as the spatial PKAM. The upsampled features are denoted as f'_t. Then, f'_t is concatenated to I_t, which is the text tattoo image. The dimension of concatenated data is $h \times w \times (c + 3)$, where h is the image height, w is the image width and c is the number of channels of f'_t. A series of convolutional layers are applied to the concatenated data to enhance the relevant regions and suppress irrelevant regions and to reduce its dimension from $h \times w \times (c + 3)$ to $h \times w \times 3$. Then, normalisation and ReLU operations are applied and obtained the enhanced image, I'_t, which is submitted to the next CNN block for extracting tattoo or text tattoo features. The convolution operations can be considered as a channel selection scheme, where the filters weight the importance of different channels.

4.2 | Network architecture

The proposed Text Tattoo Localisation Network based on Double Attention (TTLN-DA) consists of three subnetworks and two spatial PKAMs. Figure 6 gives the architecture of TTLN-DA. The three subnetworks are named as ObjectHumanNet, TattooNet and TextTattooNet. In this study, TridentNets [36] are used as the backbone of the subnetworks. TridentNet uses dilated convolutional kernels with three different dilation rates to extract the features of the target images. This architecture makes the receptive field covering objects of different sizes, from small objects to large ones. Therefore, we chose TridentNet so that we would not miss some small text tattoos or large ones in the text tattoo localisation task. Each subnetwork consists of four ResNet [49] blocks. The first three ResNet blocks consist of different number of ResNet units. The fourth block consists of three branches with different sizes of receptive field and each branch with the same number of ResNet units shares weights with the other branches. All the subnetworks are connected using spatial PKAM modules. The two spatial PKAM modules in TTLN-DA are called OPKAM and TPKAM. The OPKAM connects TattooNet to ObjectHumanNet, and TPKAM connects TextTattooNet to TattooNet. The first application of PKAM (OPKAM) is to use human instance information as attention hints to detect general tattoos. The second application of PKAM (TPKAM) is to use general tattoos as attention hints to detect text tattoos. The classification head and bounding box regression head are connected to the feature map of TextTattooNet to output the predicted class labels and bounding boxes. A multi-task loss [50] L is employed for the network training:

$$L(p, u, t^u, v) = L_{cls}(p, u) + \lambda L_{loc}(t^u, v), \tag{1}$$

in which p is the predicted class label probability distribution, $p = (p_0, p_1, \ldots, p_k)$ over $k + 1$ categories, u is the ground-truth class label, t^u is the predicted bounding box and v is the ground-truth bounding box. λ is 1 for non-background ground truth bounding boxes and 0 for background ground truth bounding boxes. L_{cls} is the classification loss and L_{loc} is the bounding box regression loss.

$$L_{cls}(p, u) = -\log p_u, \tag{2}$$

$$L_{loc}(t^u, v) = \sum_{i \in \{x,y,w,h\}} smooth_{L_1}(t_i^u - v_i), \tag{3}$$

in which p_u is the probability of the ground-truth class label in the predicted class label probability distribution and

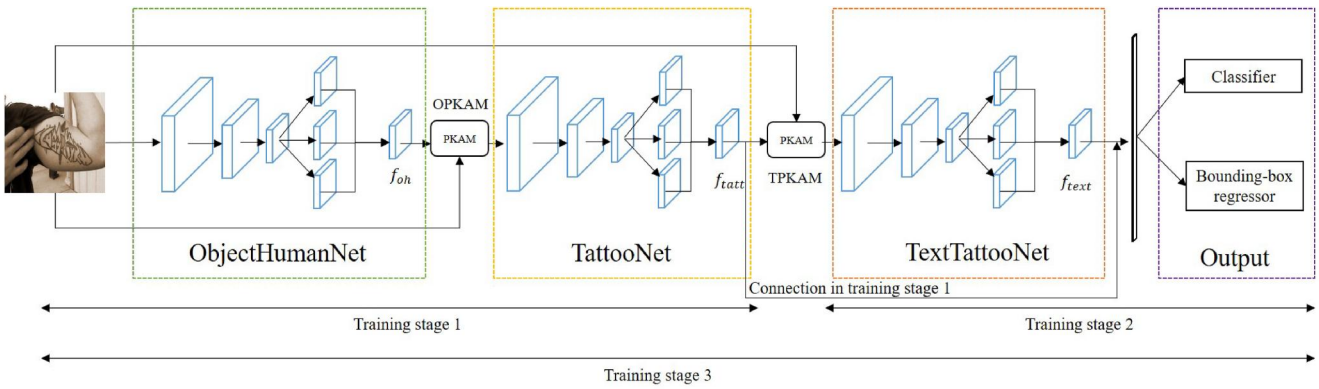

FIGURE 6 The architecture of TTLN-DA. f_{oh} is the feature map of ObjectHumanNet, f_{tatt} is the feature map of TattooNet and f_{text} is the feature map of TextTattooNet. In training stage 1, TextTattooNet and TPKAM are disconnected from the entire network, and the classification head and bounding box regression head are connected to the feature map of TattooNet

20474946, 2022, 3, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/bme2.12071, Wiley Online Library on [16/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

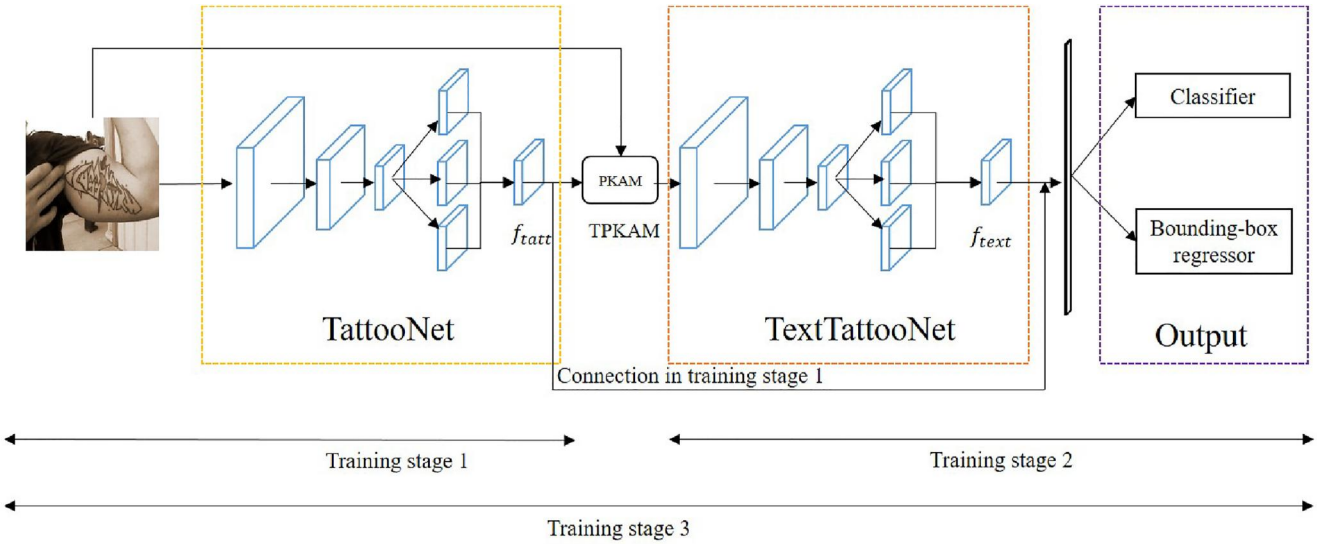

FIGURE 7 The architecture of TTLN-DA-V2. f_{tatt} is the feature map of TattooNet, and f_{text} is the feature map of TextTattooNet. In training stage 1, TextTattooNet and TPKAM are disconnected from the entire network, and the classification head and bounding box regression head are connected to the feature map of TattooNet

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2, & if\ |x| < 1 \\ |x| - 0.5, & otherwise. \end{cases} \quad (4)$$

4.3 | Variants of TTLN-DA

To seek an effective architecture, another two architectures are designed to use the dependence among objects, humans, tattoos and text tattoos in different ways. The original TTLN-DA uses two PKAM and three TridenNets to achieve double attention for text tattoo localisation. This approach makes the network very deep and large. As a result, training and testing time will be longer. Although it is not a serious problem in forensic investigation, because there is no real-time requirement, it is worth studying the necessity of using double attention. TTLN-DA-V2, which uses only two subnetworks and one spatial PKAM module to perform single attention based on the dependence between tattoos and text tattoos, is designed. Figure 7 shows its architecture. Because of the removal of the ObjectHumanNet and the OPKAM, TTLN-DA-V2's size is around 33% smaller than that of TTLN-DA.

In TTLN-DA and TTLN-DA-V2, the spatial PKAMs are inserted in-between two subnetworks. In fact, it can also be inserted within a single subnetwork and use the features within the subnetwork for attention. TTLN-DA-V3 illustrated in Figure 8 uses a channel PKAM within a single subnetwork. The channel PKAM is employed instead of spatial PKAM because of the dimension requirement of the last block in TTLN-DA-V3. This design does not increase the depth of the network significantly. In TTLN-DA-V3, the inputs of PKAM are the feature maps f_{b3} of the third block and the original image, I_t. The size of f_{b3} is $m_3 \times n_3 \times c_3$ and the size of I_t is $h \times w \times 3$. f_{b3} is upsampled to $h \times w \times c_3$ through the bilinear upsampling algorithm and then concatenated with I_t to form a feature map with a dimension of $h \times w \times (3 + c_3)$. Next, a separated convolutional operation is applied to the feature map to enhance the features and downscale its dimension to match the dimension requirement of the last block.

5 | IMPLEMENTATION AND TRAINING SCHEME

The proposed TTLN-DA and its variants are implemented in MXNet based on the SimpleDet [51] framework. The code will be available for public access [52].

5.1 | Training TTLN-DA

The proposed TTLN-DA consists of three subnetworks (ObjectHumanNet, TattooNet and TextTattooNet) and two PKAM modules (OPKAM and TPKAM). It is trained via a 3-stage training scheme. The first stage is to train the TattooNet with ObjectHumanNet fixed; the second stage is to train the TextTattooNet with both TattooNet and ObjectHumanNet fixed, and finally in the third stage, TTLN-DA is finetuned entirely.

Stage 1. In this stage, tattoos are considered as the target object and the general objects and humans are considered as relevant objects for negative and positive attention, respectively. The training set is the entire NTU Tattoo V2 dataset and the ObjectHumanNet is pretrained on MS-COCO object localisation dataset, which contains various objects and humans. In this training stage, ObjectHumanNet is fixed. When training starts, training image, I_t is inputted into ObjectHumanNet. The outputs of ObjectHumanNet are the general object and human features extracted from I_t and denoted as f_{ob}. OPKAM takes both I_t and f_{ob} as inputs and outputs the enhanced tattoo image I'_t, as shown in Figure 4, which is then inputted into TattooNet.

20474946, 2022, 3, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/bme2.12071, Wiley Online Library on [16/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

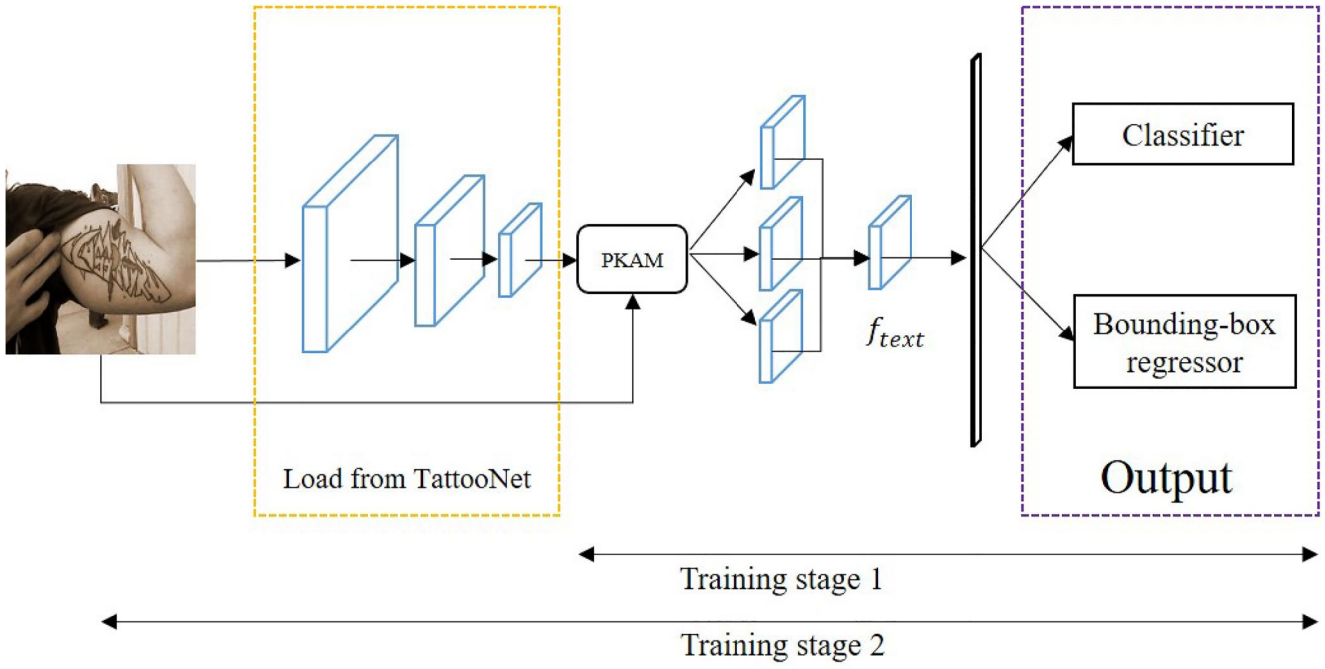

FIGURE 8 The architecture of TTLN-DA-V3. The weights of first three blocks are loaded from a pretrained TattooNet of TTLN-DA-V2. f_{text} is the final feature map

Here, the classification loss and bounding box regression loss are both applied on the last features f_{tatt} of TattooNet. TextTattooNet is disconnected from the entire network in this stage. The training is performed for six epochs and only TattooNet and OPKAM are trained in this stage.

Stage 2. In this stage, TextTattooNet is connected to TattooNet via the TPKAM and the training set is the NTU Text Tattoo V1. Text tattoo is considered as target object and tattoo is considered as a relevant object for attention. In this stage, text tattoo image, which is a training image, is denoted as I_t. All the layers in ObjectHumanNet and TattooNet are fixed. When training starts, ObjectHumanNet outputs the general object and human features f_{oh} of I_t. The OPKAM enhances I_t with its general object and human features, and the enhanced image is inputted into TattooNet, which is already trained for general tattoo localisation in the stage 1. TattooNet outputs the tattoo features of the enhanced text tattoo image, denoted as f_{tatt}. The TPKAM takes both I_t and f_{tatt} as inputs and outputs the double enhanced text tattoo image I_t''. Then, I_t'' is inputted into TextTattooNet. In this stage, the classification loss and bounding box regression loss are both applied on the last features f_{text} of TextTattooNet. The training is performed for six epochs and only TextTattooNet and TPKAM are trained in this stage.

Stage 3. In this stage, the training set is the NTU Text Tattoo V1 and entire TTLN-DA is finetuned end-to-end for additional six epochs.

The initial learning rate in the first and second stages is 0.01, while in the third stage, it is 0.0001 for ObjectHumanNet and TattooNet and 0.001 for TextTattooNet. The momentum for all layers is 0.9 throughout the training period. A pretrained MS-COCO object localisation TridentNet is used to initialise ObjectHumanNet. Xavier's algorithm [53] is used to initialise all other new layers in TTLN-DA and stochastic gradient descent (SGD) algorithm is used to train the network. The proposed TTLN-DA is trained in a batch size of 6 on 3 NVIDIA V100 GPU cards.

5.2 | Training TTLN-DA variants

TTLN-DA-V2 is trained again using a 3-stage training scheme, which is similar to the training scheme of TTLN-DA. In the first stage, TextTattooNet is disconnected from the entire network and TattooNet is trained on the entire NTU Tattoo V2 dataset from scratch for the first six epochs. The classification loss and bounding box regression loss are both applied on the last features f_{tatt} of TattooNet. In the second stage, TextTattooNet and TPKAM are trained on the NTU Text Tattoo V1 for the next six epochs with TattooNet fixed. In the final stage, the entire TTLN-DA-V2 is finetuned on the NTU Text Tattoo V1 for the last six epochs. The batch size is set to 6 for training on 3 NVIDIA V100 GPU cards. The initial learning rate for the first and second stages is 0.01 and then reduced, respectively, to 0.0001 and 0.001 for TattooNet and TextTattooNet in the third stage.

Unlike the previous two versions, TTLN-DA-V3 utilises a 2-stage training scheme. In TTLN-DA-V3, the first four blocks are loaded from a pretrained TattooNet from TTLN-DA-V2. In the first stage of training, blocks 1–3 are fixed, and only PKAM and block 4 are trained on the NTU Text Tattoo V1 for six epochs. In the second stage, the entire TTLN-DA-V3 is finetuned on the NTU Text Tattoo V1 for another six epochs. The initial learning rate in the first stage is 0.01 and in the stage 2 is 0.0001 for blocks 1–3 and 0.001 for block 4. The batch size is the same as in the previous versions.

20474946, 2022, 3, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/bme2.12071, Wiley Online Library on [16/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

6 | EXPERIMENTAL RESULTS AND DISCUSSIONS

To evaluate the performance of the proposed TTLN-DA and its variants and the importance of the additional information from general tattoos and objects, experiments were conducted on the NTU Text Tattoo V1 dataset. Experimental results of TTLN-DA and its variants are reported in Subsection A. Results of ablation studies are given in the following subsections. Subsection D and E list comparative results among TTLN-DAs, state-of-the-art object detectors and text detectors. Evaluation of general tattoo localisation is reported in Subsection F. All the experiments were conducted on three NVIDIA V100 GPU cards with 32 GB RAM each. For performance evaluation, average precision (AP) along with AP_{50}/AP_{75}, AR and F-score are used. Here AP/AR is the mean value of precisions/recalls that calculated using intersection of union (IoU) threshold from 0.50 to 0.95 with step 0.05. AP_{50}/AP_{75} is the precision calculated with IoU 0.50/0.75. F-score is calculated using equation $F\text{-}score = 2{*}AP{*}AR/(AP + AR)$. In all the experiments, the datasets were divided into training, validation and testing sets.

6.1 | Text tattoo localisation with TTLN-DA and its variants

The proposed TTLN-DA and its variants, including TTLN-DA-V2 and TTLN-DA-V3, were evaluated on the NTU Text Tattoo V1 and the evaluation results are listed in Table 3. The PKAMs in the proposed TTLN-DA and TTLN-DA-V2 are spatial PKAMs. The channel PKAM is used in TTLN-DA-V3 due to the dimension requirement of the last block in TTLN-DA-V3. The results show that TTLN-DA achieves the best performance with an AP_{50} of 0.740. TTLN-DA-V2 achieves an AP_{50} of 0.729.

In other words, the object and human attention in ObjectHumanNet offers AP_{50} improvement of 0.011. Although TTLN-DA is larger than TTLN-DA-V2, it is still a better choice because in forensic investigation, accuracy is much more important than network size and speed. TTLN-DA-V3, on the other hand, achieves a result with an AP_{50} of 0.711. Since TTLN-DA-V3 is constructed with a single TridentNet, it is suitable for fast training or limited GPU memory circumstance. Figures 9 and 10 show some localisation results from TTLN-DA. Figure 9 shows that TTLN-DA can detect text tattoos

FIGURE 9 Text tattoo localisation network based on double attention text tattoo localisation results. The red rectangle areas are enlarged on the right side of each image to show more details. The blue bounding boxes mark the localisation results. The green bounding boxes mark the ground truth. Images are resized to fill in the page

20474946, 2022, 3, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/bme2.12071, Wiley Online Library on [16/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

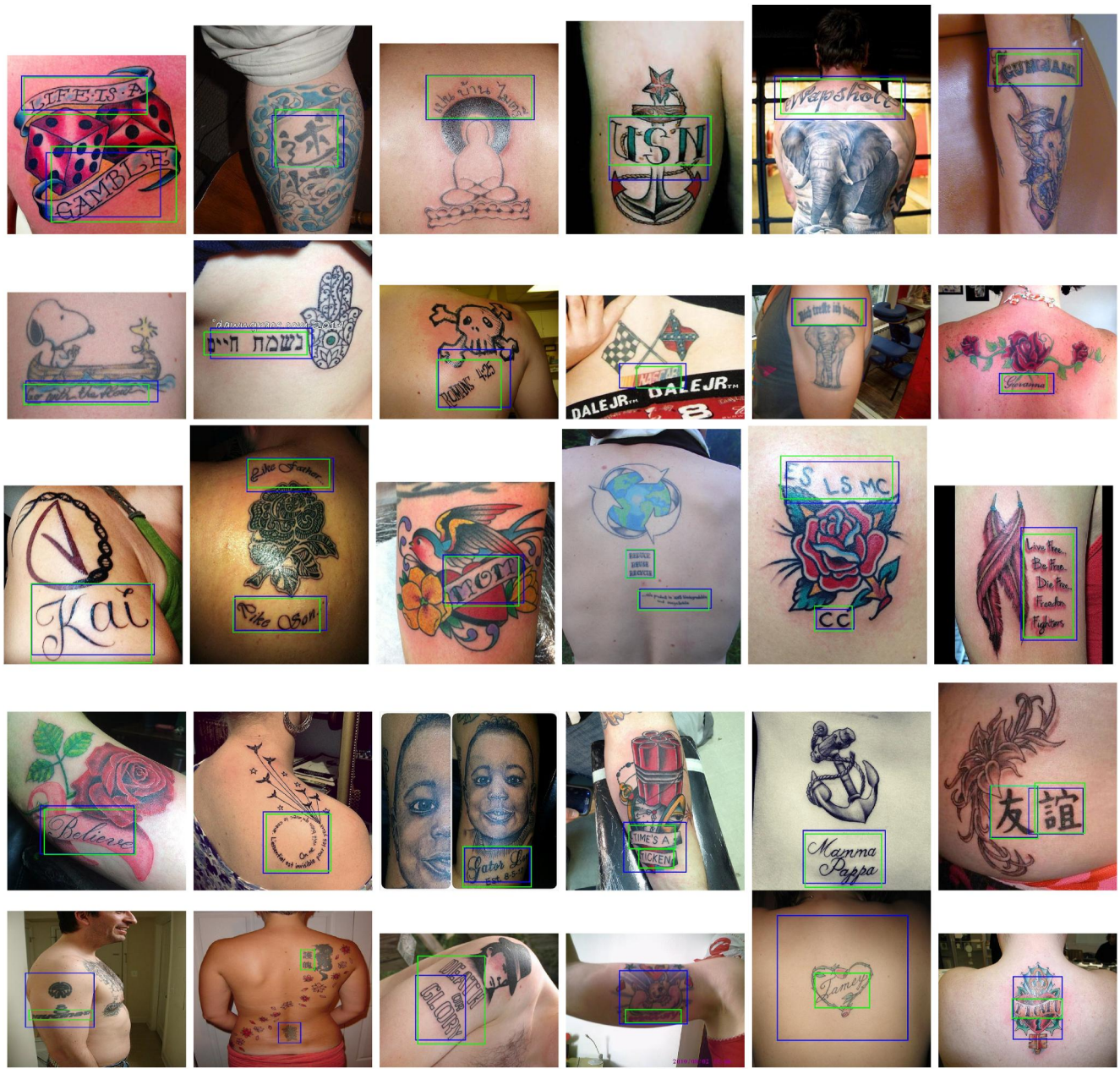

FIGURE 10 Text tattoo localisation network based on double attention (TTLN-DA) text tattoo localisation results. The blue bounding boxes mark the localisation results. The green bounding boxes mark the ground truth. Note that the first and the third images in the first row, the third and the fifth images in the second row and the second image in the fourth row are all in arbitrary shapes with various orientations; TTLN-DA can successfully detect them. The last row shows some failed localisation

Experiments	Training sets	Finetune on NTU text tattoo V1	AP_{50}
tatt-TridentNet-V1	Non-text tattoo	No	0.660
tatt-TridentNet-V2	NTU tattoo V2	No	0.684
TridentNet	NTU text tattoo V1	No	0.666
TridentNet*	NTU tattoo V2	Yes	0.689

TABLE 2 The differences between the evaluations of tatt-TridentNet-V1, tatt-TridentNet-V2, TridentNet and TridentNet*

20474946, 2022, 3, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/bme2.12071, Wiley Online Library on [16/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

in images taken in highly uncontrolled and uncooperative environments. Figure 10 shows that TTLN-DA neither wrongly detects general text nor text like tattoo patterns as text tattoos.

6.2 | Text tattoo localisation without attention

As discussed previously, the PKAM plays an important role in TTLN-DA and its variants. To examine the effectiveness of PKAMs, two networks called non-Atten-TTLN-V1 and non-atten-TTLN-V2, were constructed. Non-Atten-TTLN-V1 was constructed with three TridentNets. The TridentNets in non-Atten-TTLN-V1 were connected directly without any PKAMs. This guarantees that non-Atten-TTLN-V1 and TTLN-DA have the same network depth. Non-Atten-TTLN-V2 was constructed with only two TridentNets, which were connected directly without PKAMs. This guarantees that non-Atten-TTLN-V2 and TTLN-DA-V2 have the same network depth. Both non-Atten-TTLN-V1 and non-Atten-TTLN-V2 were trained end-to-end on the NTU Tattoo V2 dataset for 6 epochs and then finetuned on the NTU Text Tattoo V1 for 12 epochs. The total training epochs were 18, which were same as TTLN-DA and TTLN-DA-V2. The initial learning rate for non-Atten-TTLN-V1 and non-Atten-TTLN-V2 was 0.01. The results listed in Table 3 show that non-Atten-TTLN-V1 and non-Atten-TTLN-V2 perform worse than TTLN-DA and TTLN-DA-V2, which have the same network depths, respectively. Their performance differences are very significant. Without the PKAMs, non-Atten-TTLN-V1 and non-Atten-TTLN-V2 can easily suffer from over-fitting because of their depths. TTLN-DA and TTLN-DA-V2 are able to avoid the over-fitting problem because the PKAMs highlight relevant features and suppress irrelevant features, such that learning in deep layers becomes easier.

6.3 | Text tattoo localisation with general tattoo detector

Some might argue that a well-trained non-text tattoo detector should be able to detect text tattoos effectively, since text tattoo is a subcategory of tattoos. To evaluate whether general tattoo features are sufficient for performing text tattoo localisation, an experiment was conducted with a single TridentNet, which was trained on the non-text tattoo images in the NTU Tattoo V2 and tested on the NTU Text Tattoo V1. This TridentNet is called tatt-TridentNet-V1. Another experiment was conducted with a single TridentNet, which was trained on the NTU Tattoo V2 but without fine-tuning on the NTU Text Tattoo V1. As with tatt-TridentNet-V1, it was tested on the NTU Text Tattoo V1. This TridentNet is called tatt-TridentNet-V2. The initial learning rate for training tatt-TridentNet-V1 and tatt-TridentNet-V2 was 0.01, and the batch size was six on the GPUs. The evaluation results reported in Table 3 show that tatt-TTLN-V1 performs worse than tatt-TTLN-V2, because tatt-TTLN-V2 was trained on the NTU Tattoo V2 containing not only non-text tattoo images but also text tattoo images. This result indicates that non-text tattoo features cannot fully represent text tattoo images. Tatt-Trident-V2 performs better than the TridentNet in Table 3, which was trained only on the NTU Text Tattoo V1. This result pinpoints that non-text tattoo images can improve text tattoo

TABLE 3 Experimental results on text tattoo localisation

Experiments	AP	AP_{50}	AP_{75}	AR	F-score
TTLN-DA	**0.393**	**0.740**	**0.380**	**0.579**	**0.468**
TTLN-DA-V2	0.382	0.729	0.356	0.575	0.459
TTLN-DA-V3	0.370	0.711	0.343	0.576	0.451
Non-atten-TTLN-V1	0.187	0.336	0.154	0.423	0.259
Non-atten-TTLN-V2	0.239	0.428	0.223	0.403	0.300
tatt-TridentNet-V1	0.334	0.660	0.305	0.528	0.409
tatt-TridentNet-V2	0.342	0.684	0.305	0.550	0.422
TridentNet	0.323	0.666	0.279	0.521	0.399
TridentNet*	0.350	0.689	0.315	0.553	0.427
CSPNet	0.239	0.596	0.128	0.231	0.235
CSPNet*	0.326	0.681	0.280	0.310	0.318
DetectoRS	0.312	0.652	0.267	0.484	0.379
DetectoRS*	0.340	0.679	0.288	0.524	0.412
Cascade R-CNN	0.321	0.615	0.287	0.507	0.393
Cascade R-CNN*	0.349	0.660	0.324	0.528	0.420
CenterNet	0.317	0.623	0.274	0.492	0.386
CenterNet*	0.335	0.655	0.314	0.530	0.411
RetinaNet	0.298	0.591	0.259	0.476	0.367
RetinaNet*	0.315	0.645	0.297	0.509	0.389
Mask R-CNN	0.302	0.598	0.264	0.481	0.371
Mask R-CNN*	0.324	0.643	0.273	0.511	0.397
Faster R-CNN	0.284	0.587	0.244	0.457	0.350
Faster R-CNN*	0.303	0.617	0.258	0.472	0.369
EAST	0.304	0.608	0.274	0.162	0.211
SegLink	0.188	0.462	0.127	0.140	0.160
PixelLink	0.148	0.321	0.117	0.201	0.170
PAN	0.314	0.641	0.286	0.384	0.345
PSENet	0.171	0.377	0.142	0.170	0.171
SPCNet	0.043	0.118	0.023	0.591	0.080
TextFuseNet	0.331	0.649	0.307	0.470	0.388
ABCNet	0.327	0.680	0.280	0.076	0.123
DCINN	0.320	0.630	0.292	0.548	0.404

Note: * mark means that the networks are trained on the NTU Tattoo V2 first and finetuned on the NTU Text Tattoo V1. The bold values indicate the best results.

Abbreviations: AP, average precision; CSPNet, cross stage partial network; DCINN, deformable convolution and inception; PAN, Pixel Aggregation Network; PSENet, progressive scale expansion network; TTLN-DA, text tattoo localisation network based on double attention.

20474946, 2022, 3, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/bme2.12071, Wiley Online Library on [16/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

localisation performance. However, tatt-TridentNet-V2 still performs slightly worse than TridentNet*, which was trained on the NTU Tattoo V2 and finetuned on the NTU Text Tattoo V1, because tatt-TridentNet-V2 trained on the NTU Tattoo V2 mixes the features of non-text tattoo and text tattoo and lowers down the text tattoo localisation accuracy. It indicates that the non-text tattoo features and text tattoo features are not the same. Thus, text tattoos should not be regarded as general tattoos in localisation. Table 2 summarises the differences between the evaluations of tatt-TridentNet-V1, tatt-TridentNet-V2, TridentNet and TridentNet*.

6.4 | Comparison with state-of-the-art object detectors

TTLN-DA and its variants are compared with the state-of-the-art object detectors, including Faster R-CNN [10], Mask R-CNN [31], Cascade R-CNN [54], CenterNet [55], RetinaNet [56], CSPNet [37], DetectoRS [38] and TridentNet [36]. The backbone networks of TridentNet, Detectors, RetinaNet, Mask R-CNN and Faster R-CNN are ResNet-101; the backbone network of CSPNet is ResNeXt and the backbone network of CenterNet is Hourglass-104. The comparison results are reported in Table 3. To make a fair comparison, each method was evaluated twice. In the first evaluation, the NTU Text Tattoo V1 was used for both training and testing and all the methods were trained for six epochs to make sure that the training of all those methods are converged. In the second evaluation, the methods were trained on the NTU Tattoo V2 dataset for six epochs first, then were finetuned on the NTU Text Tattoo V1 for another six epochs. The initial learning rate was 0.01 in all the training. The batch size was six on the GPUs. The comparison results show that among the state-of-the-art object detectors, TridentNet achieves the best performance in terms of AP_{50}, which is the reason why TridentNet is used as the subnetwork in TTLN-DA and its variants. The results also show that TTLN-DA, TTLN-DA-V2

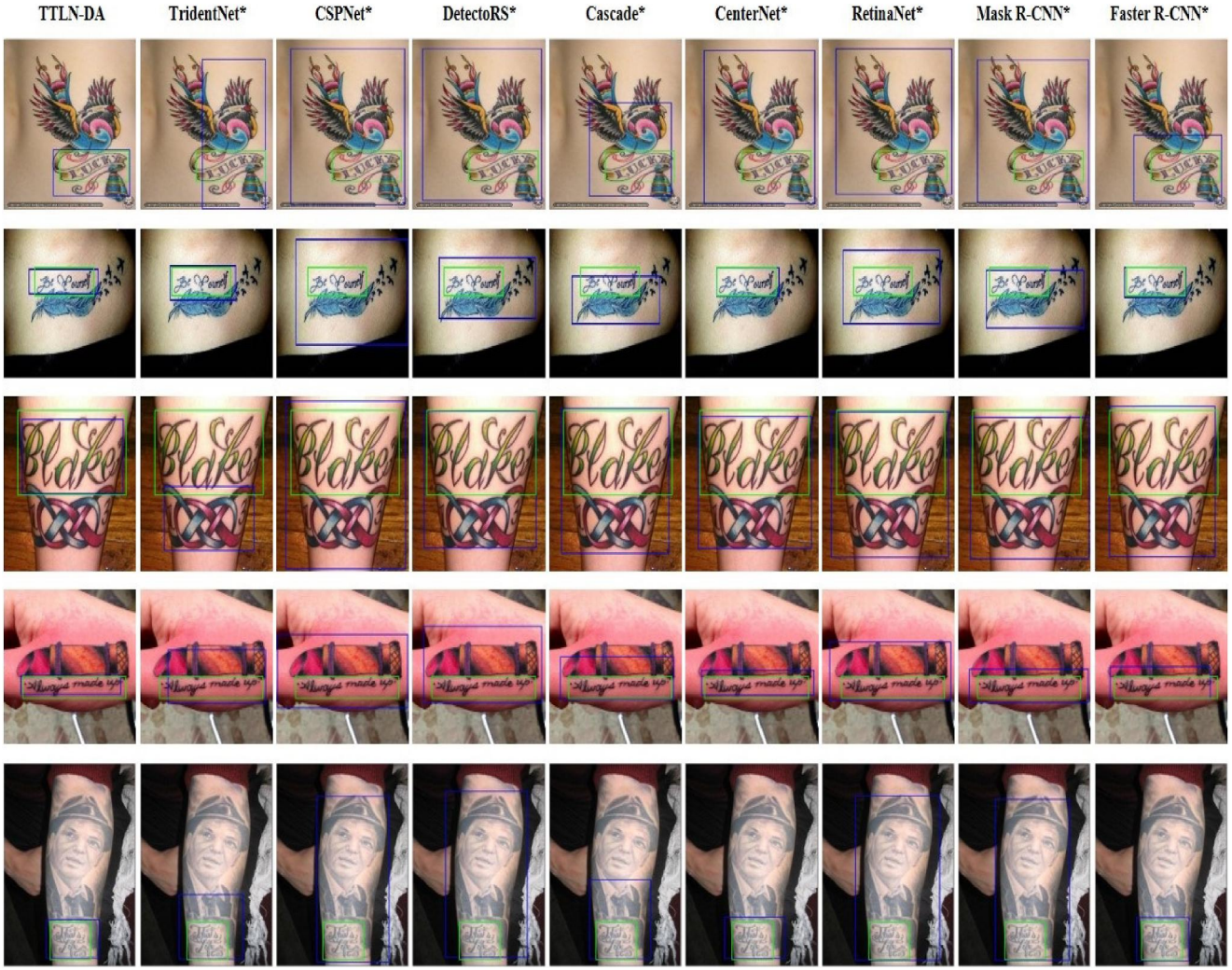

FIGURE 11 Text tattoo localisation results from close up images for comparisons. From left to right: text tattoo localisation network based on double attention, TridentNet*, cross stage partial network*, Detectors*, Cascade R-CNN*, CenterNet*, RetinaNet* Mask R-CNN* and Faster R-CNN*. The blue bounding boxes mark the localisation results. The green bounding boxes mark the ground truth

20474946, 2022, 3, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/bme2.12071, Wiley Online Library on [16/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and TTLN-DA-V3 outperform TridentNet. It is because the PKAMs provide additional attention information to the detectors. Figure 11 gives some localisation results from TTLN-DA and the state-of-the-art object detectors. The results show that TTLN-DA is able to detect text tattoos without influenced by other tattoos. Other methods detect general tattoos as well as text tattoos. TTLN-DA can also handle different text tattoo styles and text tattoo colours (see Figures 9 and 10).

6.5 | Comparison with state-of-the-art text detectors

In addition to object detectors, TTLN-DA and its variants, including TTLN-DA-V2 and TTLN-DA-V3, are also compared with state-of-the-art text detectors, including EAST [25], SegLink [26], PixelLink [27], PAN [28], PSENet [32], SPCNet [29], TextFuseNet [34] and ABCNet [35]. The backbone networks of EAST, PAN, PSENet, SPCNet, TextFuseNet and ABCNet are ResNet-101. The backbone networks of SegLink and PixelLink are VGG-16. All these detectors were pretrained text localisation models and finetuned on the NTU Text Tattoo V1. Note that we also compare with DCINN [15], which was proposed for text tattoo localisation. Since the authors of DCINN neither released their source code nor dataset, we have to re-implement the DCINN based on their description of the network architecture. The software version and training parameters might not be exactly the same. Thus, the results of DCINN are reported here only as a reference. Table 3 shows that some state-of-the-art text detectors cannot handle text tattoo localisation well enough especially for images with non-text tattoos (see Figure 12). These results indicate that text tattoos are different significantly from general texts in images. These differences have been discussed in the introduction and illustrated in Figure 3.

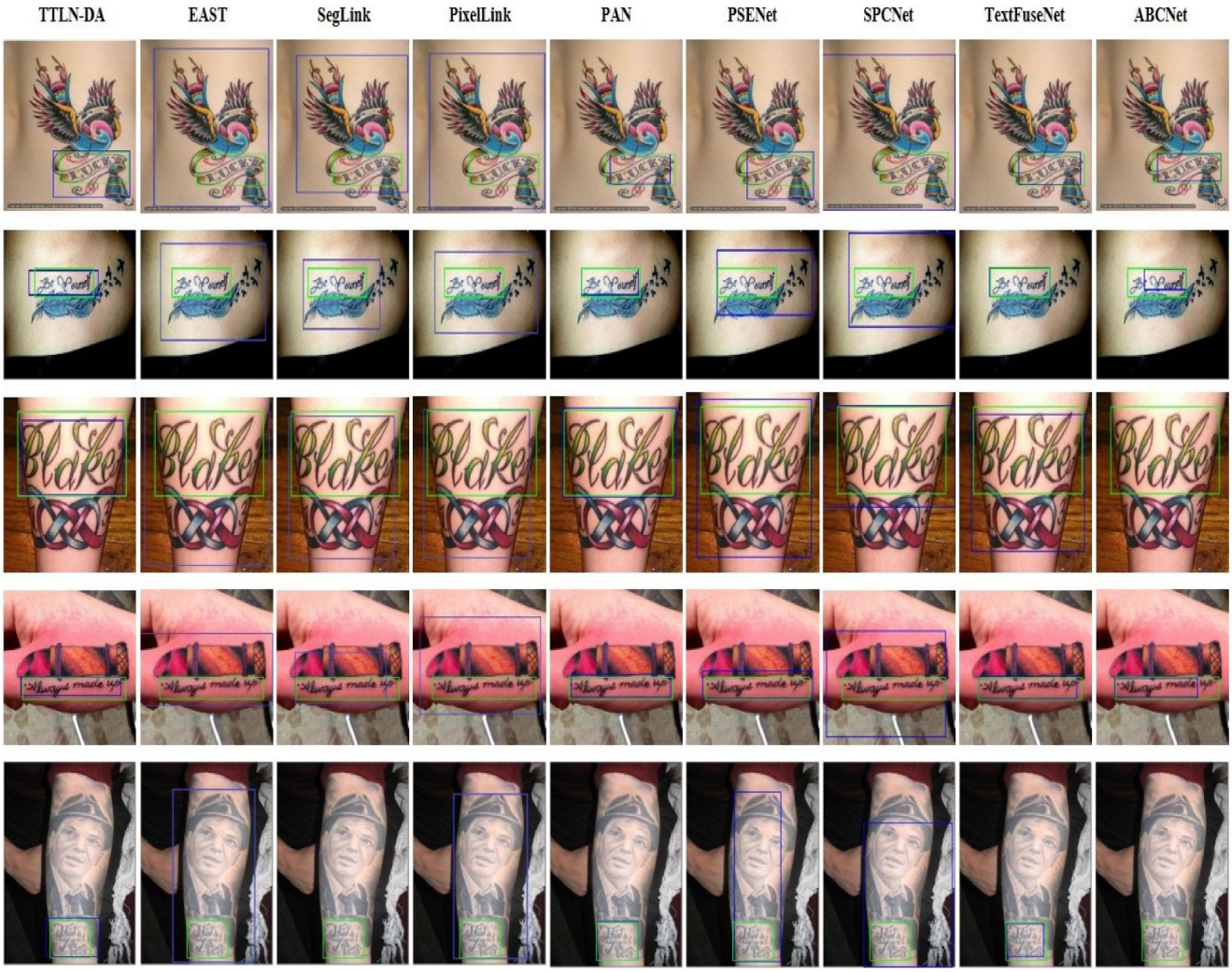

FIGURE 12 Text tattoo localisation results from close up images for comparisons. From left to right: text tattoo localisation network based on double attention, EAST, PixelLink, SegLink, pixel aggregation network, progressive scale expansion network, SPCNet, TextFuseNet and ABCNet. The blue bounding boxes mark the localisation results. The green bounding boxes mark the ground truth

20474946, 2022, 3, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/bme2.12071, Wiley Online Library on [16/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

6.6 | Evaluation of general tattoo localisation

TTLN-DA and TTLN-DA-V2 utilise general tattoo features as attention to improve the text tattoo localisation performance. They are also able to perform general tattoo localisation by using the outputs of TattooNet. To evaluate their performance for general tattoo localisation, experiments were conducted. As the previous experiments, the state-of-the-art detectors are employed for comparisons. These object detectors were trained on the NTU Tattoo V2. Table 4 lists the experimental results.

In TTLN-DA, TattooNet is the second subnetwork that follows the ObjectHumanNet. The ObjectHumanNet provides the general object features, which are used as the attention in the OPKAM, and TattooNet takes the enhanced general tattoo images as training data. Therefore, TTLN-DA outperforms TTLN-DA-V2 and the state-of-the-art object detectors. TTLN-DA-V2 performs similar to TridentNet because it does not use the object features as attention for tattoo localisation and its tattoo localisation network is the same as TridentNet.

TABLE 4 Experimental results on the general tattoo localisation

Experiments	AP	AP_{50}	AP_{75}	AR	F-score
TTLN-DA	**0.445**	**0.807**	**0.433**	**0.587**	**0.506**
TTLN-DA-V2	0.414	0.770	0.390	0.565	0.478
TridentNet*	0.416	0.760	0.403	0.571	0.481
CSPNet*	0.371	0.735	0.334	0.329	0.349
DetectoRS*	0.388	0.733	0.366	0.533	0.449
Cascade R-CNN*	0.396	0.738	0.375	0.533	0.454
CenterNet*	0.407	0.774	0.382	0.552	0.469
RetinaNet*	0.344	0.686	0.304	0.544	0.421
Mask R-CNN*	0.373	0.758	0.344	0.533	0.439
Faster R-CNN*	0.393	0.753	0.363	0.520	0.448

Note: NTTLN-DA and TTLN-DA-V2 use the outputs of TattooNet as the general tattoo localisation results. The bold values indicate the best results.

Abbreviations: AP, average precision; CSPNet, cross stage partial network; TTLN-DA, text tattoo localisation network based on double attention.

7 | CONCLUSION AND FUTURE WORK

The previous tattoo localisation studies applied general object detectors on small scale tattoo datasets. Text tattoos, which have rich personal information, were totally neglected. For studying text tattoo localisation, two new datasets, the NTU Tattoo V2 and the NTU Text Tattoo V1, are established. The NTU Tattoo V2 is the largest tattoo dataset and the NTU Text Tattoo V1 is the only text tattoo dataset. For using the dependence among objects, humans, tattoos and text tattoos, the prior knowledge-based attention mechanism (PKAM) is proposed. Three different network architectures with PKAM are designed and examined. The architecture with double attention from objects, humans and tattoos for text tattoo localisation, named Text Tattoo Localisation Network based on Double Attention (TTLN-DA), achieves the best results among the three architectures. Evaluation results showed that TTLN-DA can detect text tattoos in arbitrary shapes with various orientations. Extensive experiments were performed. The results indicate the importance of the dependence and effectiveness of PKAM. They also demonstrate that TTLN-DA and its variants outperform the state-of-the-art object detectors and text detectors for both tattoo and text tattoo localisation. Networks with same architecture but trained on different datasets were also evaluated and the evaluation results pinpointed that non-text tattoo images can improve text tattoo localisation performance.

Text tattoo localisation is the first step towards text tattoo recognition, which will be a new forensic tool for processing text tattoo images. With this tool, law enforcement agencies

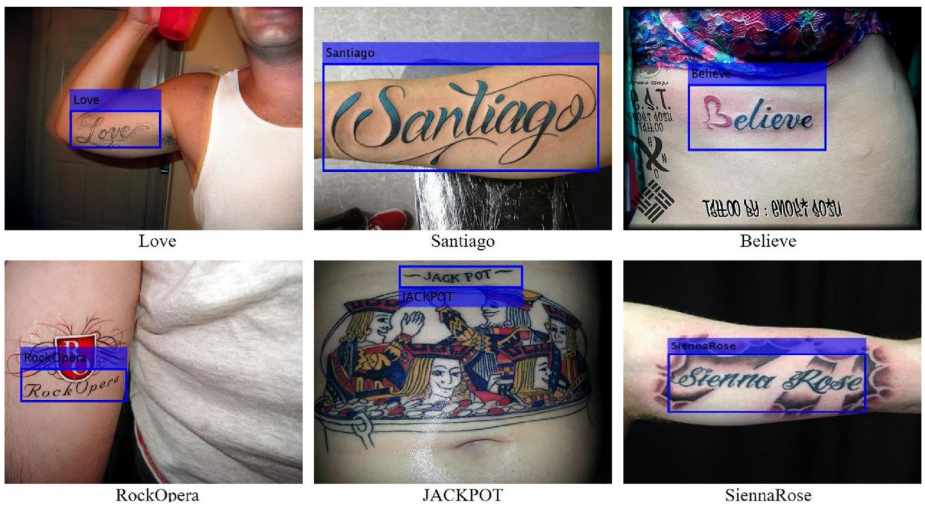

FIGURE 13 Sample images of text tattoo recognition. The blue bounding box marks the localised text tattoos using text tattoo localisation network based on double attention. The recognised texts are given below the images

20474946, 2022, 3, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/bme2.12071, Wiley Online Library on [16/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

FIGURE 14 Comparison of recognition results based on three different localisation results. From left to right: text tattoo localisation network based on double attention (TTLN-DA), TridentNet* and ABCNet. The blue bounding box marks the localised text tattoos using TTLN-DA. The recognised texts are given below the images

can employ text-based tattoo searching for forensic investigations. Some preliminary experiments on recognising tattoo text have been performed. Applying an existing scene text recognition method [35] on localised text tattoo images from TTLN-DA, TridentNet* and ABCNet, text can be recognised if the text tattoos are well localised. Figure 13 shows some text tattoo recognition results. Figure 14 demonstrates that the text tattoo recognition results are affected by the localisation quality. More efforts will be put on text tattoo recognition task in the future.

ACKNOWLEDGEMENT

This work is partially supported by the Ministry of Education, Singapore through Academic Research Fund Tier 1, RG21/19-(S).

CONFLICT OF INTEREST

This manuscript has not been published and is not under consideration for publication elsewhere. We have no conflicts of interest to disclose.

DATA AVAILABILITY STATEMENT

The data that support the findings will be available at https://github.com/BFLTeam/NTU Dataset from the date of publication to allow for academic research findings.

ORCID

Xingpeng Xu https://orcid.org/0000-0002-0791-0221
Kuanhong Cheng https://orcid.org/0000-0002-0025-4890

END NOTE

[1] https://github.com/BFLTeam/NTU_Dataset.

REFERENCES

1. Economist, T.: A statistical analysis of the art on convicts' bodies. https://www.economist.com/christmas-specials/2016/12/24/a-statistical-analysis-of-the-art-on-convicts-bodies
2. History of Tattoos: Tattoo statistics—how many people have tattoos? http://www.historyoftattoos.net/tattoo-facts/tattoo-statistics/
3. Di, X., Patel, V.M.: Deep tattoo recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 119–126. (2016)
4. Ngan, M., Grother, P.: Tattoo recognition technology—challenge (Tatt-c): an open tattoo database for developing tattoo recognition research. In: IEEE International Conference on Identity, Security and Behavior Analysis (ISBA), pp. 1–6. (2015)
5. Ngan, M., Quinn, G.W., Grother, P.J.: Tattoo recognition technology - challenge (Tatt-c): outcomes and recommendations (2015)
6. Shan, S., et al.: Robust Tattoo Detection and Retrieval Competition (RTDRC) (2019). https://sites.google.com/site/rtdrc2019/
7. Huynh, N.Q., et al.: A preliminary report on a full-body imaging system for effectively collecting and processing biometric traits of prisoners. In: IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM), pp. 167–174. (2014)
8. Hrkac, T., et al.: Deep learning architectures for tattoo detection and de-identification. In: First International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE), pp. 1–5. (2016)
9. Sun, Z., et al.: Tattoo detection and localization using region-based deep learning. In: 23rd International Conference on Pattern Recognition (ICPR), pp. 3055–3060. (2016)
10. Ren, S., et al.: Faster r-CNN: towards real-time object detection with region proposal networks. Adv. Neural Inf. Process Syst. 39, 1137–1149. (2015)
11. Heflin, B., Scheirer, W.J., Boult, T.E.: Detecting and classifying scars, marks, and tattoos found in the wild. In: IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 31–38. (2012)
12. Schölkopf, B., Platt, J., Hofmann, T.: Graph-based visual saliency. In: Neural Information Processing Systems (NIPS), pp. 545–552. (2006)
13. Rother, C., Kolmogorov, V., Blake, A.: Grabcut-interactive foreground extraction using iterated graph cuts. ACM Trans. Graph 23(3), 309–314. (2004)
14. Han, H., et al.: Tattoo image search at scale: joint detection and compact representation learning. IEEE Trans. Pattern Anal. Mach. Intell. 41, 2333–2348. (2018)
15. Chowdhury, T., et al.: DCINN: deformable convolution and inception based neural network for tattoo text detection through skin region. In: International Conference on Document Analysis and Recognition, pp. 335–350. (2021)
16. Epshtein, B., Ofek, E., Wexler, Y.: Detecting text in natural scenes with stroke width transform. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2963–2970. (2010)
17. Neumann, L., Matas, J.: Real-time scene text localization and recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3538–3545. (2012)
18. Anthimopoulos, M., Gatos, B., Pratikakis, I.: Detection of artificial and scene text in images and video frames. Pattern Anal. Appl. 16, 431–446. (2013)
19. Chen, H., et al.: Robust text detection in natural images with edge-enhanced maximally stable extremal regions. In: IEEE Conference on Computer Vision and Pattern Recognition (ICIP), pp. 2609–2612. (2011)
20. Posner, I., Corke, P.I., Newman, P.: Using text-spotting to query the world. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3181–3186. (2010)

20474946, 2022, 3, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/bme2.12071, Wiley Online Library on [16/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

21. Wang, K., Babenko, B., Belongie, S.J.: End-to-end scene text recognition. In: International Conference on Computer Vision (ICCV), pp. 1457–1464. (2011)
22. Rong, X., Yi, C., Tian, Y.: Unambiguous text localization and retrieval for cluttered scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3279–3287. (2017)
23. Zhong, Z., et al.: Deeptext: a unified framework for text proposal generation and text detection in natural images. ArXiv. arXiv preprint arXiv:1605.07314 (2016)
24. Liao, M., et al.: Textboxes: a fast text detector with a single deep neural network. In: AAAI Conference on Artificial Intelligence (AAAI) (2016)
25. Zhou, X., et al.: East: an efficient and accurate scene text detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2642–2651. (2017)
26. Shi, B., Bai, X., Belongie, S.J.: Detecting oriented text in natural images by linking segments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3482–3490. (2017)
27. Deng, D., et al.: Pixellink: detecting scene text via instance segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), vol. 32, no. 1 (2018)
28. Wang, W., et al.: Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8439–8448. (2019)
29. Xie, E., et al.: Scene text detection with supervised pyramid context network. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 33, pp. 9038–9045. (2019)
30. Lin, T.-Y., et al.: Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 936–944. (2017)
31. He, K., et al.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988. (2017)
32. Wang, W., et al.: Shape robust text detection with progressive scale expansion network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9336–9345. (2019)
33. Liao, M., et al.: Real-time scene text detection with differentiable binarization. In: Proceedings of the AAAI Conference on Artificial Intelligence, 11474–11481. (2020)
34. Ye, J., et al.: Textfusenet: scene text detection with richer fused features. In: IJCAI. vol. 20, pp. 516–522. (2020)
35. Liu, Y., et al.: Abcnet: real-time scene text spotting with adaptive bezier-curve network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9806–9815. (2020)
36. Li, Y., et al.: Scale-aware trident networks for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6054–6063. (2019)
37. Wang, C.-Y., et al.: Cspnet: a new backbone that can enhance learning capability of CNN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1571–1580. (2020)
38. Qiao, S., Chen, L.-C., Yuille, A.: Detectors: detecting objects with recursive feature pyramid and switchable atrous convolution. arXiv, preprint arXiv:2006.02334 (2020)
39. Hrkac, T., Brkic, K., Kalafatic, Z.: Tattoo detection for soft biometric de-identification based on convolutional neural networks. In: Proceedings of the OAGM-ARW Joint Workshop (OeAGM/AAPR), pp. 131–138. (2016)
40. Xu, Q., et al.: Tattoo detection based on cnn and remarks on the nist database. In: International Conference on Biometrics (ICB), pp. 1–7. (2016)
41. Mnih, V., et al.: Recurrent models of visual attention. In: Neural Information Processing Systems (NIPS). vol. 27 (2014)
42. Ba, J., Mnih, V., Kavukcuoglu, K.: Multiple object recognition with visual attention. arXiv. preprint arXiv:1412.7755 (2014)
43. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning (ICML), pp. 2048–2057. (2015)
44. Gregor, K., et al.: Draw: a recurrent neural network for image generation. In: International Conference on Machine Learning (ICML), pp. 1462–1471. (2015)
45. Jaderberg, M., et al.: Spatial transformer networks. Adv. Neural Inf. Process. Syst. 28 (2015)
46. Almahairi, A., et al.: Dynamic capacity networks. In: International Conference on Machine Learning (ICML) (2015)
47. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141. (2017)
48. Woo, S., et al.: Cbam: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19. (2018)
49. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. (2015)
50. Girshick, R.B.: Fast R-CNN. In: ICCV, pp. 1440–1448. (2015)
51. Chen, Y., et al.: Simpledet: a simple and versatile distributed framework for object detection and instance recognition. J. Mach. Learn. 20(156), 1–8. (2019)
52. NTU: Ntu Forensic Image Databases (2020). https://github.com/BFLTeam/NTU_Dataset. Accessed 03 01
53. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: AISTATS (2010)
54. Cai, Z., Vasconcelos, N.: Cascade R-CNN: delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6154–6162. (2018)
55. Duan, K., et al.: Centernet: keypoint triplets for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6569–6578. (2019)
56. Lin, T.-Y., et al.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2999–3007. (2017)

How to cite this article: Xu, X., et al.: Using double attention for text tattoo localisation. IET Biome. 11(3), 199–214 (2022). https://doi.org/10.1049/bme2.12071

20474946, 2022, 3, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/bme2.12071, Wiley Online Library on [16/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License