This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3378458

TTS: Hilbert Transform-based Generative Adversarial Network for Tattoo and Scene Text Spotting

Ayan Banerjee, Shivakumara Palaiahnakote, *Senior Member IEEE*, Umapada Pal, *Senior Member IEEE*, Apostolos Antonacopoulos, *Member IEEE*, Tong Lu and Josep Llados Canet

***Abstract*— Text spotting in natural scenes is of increasing interest and significance due to its critical role in several applications, such as visual question answering, named entity recognition and event rumor detection on social media. One of the newly emerging challenging problems is Tattoo Text Spotting (TTS) in images for assisting forensic teams and for person identification. Unlike the generally simpler scene text addressed by current state-of-the-art methods, tattoo text is typically characterized by the presence of decorative backgrounds, calligraphic handwriting and several distortions due to the deformable nature of the skin. This paper describes the first approach to address TTS in a real-world application context by designing an end-to-end text spotting method employing a Hilbert transform-based Generative Adversarial Network (GAN). To reduce the complexity of the TTS task, the proposed approach first detects fine details in the image using the Hilbert transform and the Optimum Phase Congruency (OPC). To overcome the challenges of only having a relatively small number of training samples, a GAN is then used for generating suitable text samples and descriptors for text spotting (i.e. both detection and recognition). The superior performance of the proposed TTS approach, for both tattoo and general scene text, over the state-of-the-art methods is demonstrated on a new TTS-specific dataset (publicly available[1]) as well as on the existing benchmark natural scene text datasets: Total-Text, CTW1500 and ICDAR 2015.**

***Index Terms*— Hilbert transform, Text detection, Text spotting, Generative adversarial networks, Calligraphic text, Tattoo text spotting.**

I. INTRODUCTION

Text spotting in natural scene images is receiving special attention because it can serve several real-time applications, such as visual question answering [1], named entity recognition [2], and event rumor detection [3] on social media platforms. This has resulted in models [1, 4] aiming to overcome challenges such as arbitrarily oriented text, irregularly shaped text, text in multiple scripts, and dense text.

Increasingly more challenging applications have also been appearing that involve text spotting for person identification and tracking e.g., in marathons and other sports [5, 6]. Those methods [5, 6] attempt to detect and recognize single characters or digits, text on clothes deformed by movement, and partially occluded text.

Ultimately the goal is to create a generalizable method widely applicable to real-world problems (simple and complex text situations). This trend has motivated the authors to create a new text spotting approach that solves a particularly challenging problem while also demonstrating excellent performance in general situations.

The challenge addressed in this work is the very recently emerging problem of tattoo *text* spotting (TTS). General tattoo *image* detection and recognition is not a new problem for the computer vision and image processing community [7]. A key objective of such methods is to assist forensic teams in identifying a person, a crime or a gang.

The underlying reason is that since each tattoo symbol or drawing is bespoke, varying uniquely according to an individual's creativity and expression, an automated method can assist a forensic team in obtaining clues about a crime, a gang, or a person. In particular, a broad observation is that drawing tattoo *text* on human body parts nowadays is fashionable, especially among celebrities [7].

PAN++ [1] Huang et al. [4] Proposed

PAN++ [1] Huang et al. [4] Proposed

Fig. 1. Text spotting by the state-of-the-art and the proposed approaches in full images (top: tattoo text, bottom: scene text).

Similarly, the main intention of having *text* tattooed on human body parts is to express individuality, one's views, convey a message etc. This is very important in several real-world applications such as person identification, personality assessment and psychological evaluation. For instance, tattoo text in images uploaded on social media can help to gain further insights on the corresponding people. Therefore, the proposed approach focuses on tattoo *text* spotting rather than drawings or symbols. After all, the text provides richer semantic information than drawings. To the best of the authors' knowledge, this is the first work on Tattoo Text Spotting (TTS)

- Ayan Banerjee and Umapada Pal are with the Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India. Email: ab2141@cse.jgec.ac.in, umapada@isical.ac.in.
- Shivakumara Palaiahnakote and Apostolos Antonacopoulos are with the School of Science, Engineering and Environment, University of Salford, Salford, M5 4WT, UK, E-mail: S.Palaiahnakote@salford.ac.uk. A.Antonacopoulos@primaresearch.org.
- Tong Lu is with the National Key Lab for Novel Software Technology, Nanjing University, Nanjing, China, E-mail: lutong@nju.edu.cn.
- Josep Llados Canet is with the Computer Vision Center, Universitat Autònoma de Barcelona, Spain, josep@cvc.uab.es

[1]https://drive.google.com/drive/folders/1o4WYa0gXuFWx6hlEGnpaZV-sEz8zElB4?usp=sharing

addressing the challenges of real-world applications.

The main challenges of TTS compared to other text spotting applications are the presence of significant variations due to the freestyle handwriting, dense calligraphic text, and complex backgrounds due to decorative features and the nature of skin itself. State-of-the-art methods on text spotting [1, 4] have not been designed with such complexities in mind and, therefore, are not effective in those situations. It is evident from Fig. 1, that the state-of-the-art methods [1, 4], which employ a kernel-based network and a transformer, respectively, for text spotting in scene images, do not perform well for the representative tattoo text image example (top row) nor for the similarly complex text in a scene image (bottom row). As mentioned earlier, this is due to the limitations and relatively narrow scope of the existing methods. On the other hand, the method proposed in this paper performs well for both the tattoo and the scene text in the example images.

While it is evident from the above that TTS is an open challenge on its own, this paper proposes a novel end-to-end approach for text spotting in both tattoo *and* natural scene images. To detect fine details (edges) in images of either type, especially given the need to overcome the varying quality and complex nature of tattoo text, the proposed method first employs the Hilbert transform (HT) and Optimum Phase Congruency (OPC) [8]. Next, the strong discriminative power of the Generative Adversarial Network (GAN) and its capabilities for synthesizing images using different features [9, 10] are employed in a novel combination to *detect and recognize* (i.e. spot) both scene and tattoo text effectively, completing the end-to-end approach.

The following are the key contributions. (i) The use of the Hilbert Transform for detecting the fine details (edges) which represent text information irrespective of whether it is a tattoo or a scene image. Since this step reduces background complexity, the performance of text spotting improves for tattoo text as well as natural scene text. (ii) The exploitation of a GAN architecture within the system to generate possible synthetic text variants based on a few original samples in order to reduce dependency on a large number of labelled samples. This is necessary especially for tattoo text images (difficult to acquire large datasets) and in turn it improves the proposed method's generalization ability. (iii) The proposed method successfully integrates the HT and GAN in an end-to-end system in a novel way for achieving superior results for both tattoo and scene text images without requiring large training datasets (for the case of TTS, where data is scarce) and without additional computational overhead compared to the state-of-the-art. (iv) A newly created Tattoo Text Spotting Dataset (TTSD) to support this newly introduced tattoo text spotting challenge for forensic and security applications.

The rest of the paper is organized as follows. Section II presents a review of existing text detection, recognition and spotting methods. The proposed approach and its use of the Hilbert transform to detect fine details in the input image and the GAN for Tattoo Text Spotting are detailed in Section III. Experimental analysis and results on both a new tattoo text dataset as well as on standard benchmark natural scene text datasets are reported and discussed in Section IV. The main findings are summarized, and future work is proposed in Section V.

II. Background and Literature Review

Since the proposed approach is a text spotting (i.e. detection and recognition) system for both tattoo and scene text, the closest state-of-the-art methods reviewed below are for text detection, recognition, and end-to-end spotting methods for natural scenes.

A. Approaches for Text Detection or Text Recognition in Natural Scenes

Deng et al. [11] proposed an efficient scene text detection method based on a recurrent feature refinement network. Raisi et al. [12] developed a model for text detection in the wild using a transformer-based network. Chowdhury et al. [5] introduced an episodic learning network for text detection in sports scene images. It does not, however, consider the detection of tattoo text images for detection. Similarly, several deep learning-based models have been developed to address the challenges of text detection in natural scene images [13-25]. However, the scope of those methods is limited to regular scene text recognition and does not include tattoo text images. Recently, Nandanwar et al. [26] proposed a model for text detection in 3D video based on the combination of a wavefront and a deep learning approach. This approach does not work well for tattoo text images because of the use of elaborately decorated characters and backgrounds. Overall, none of the above approaches considers tattoo images, and the scene text they address does not include calligraphic text like that in tattoos.

The same conclusions can be drawn for the state-of-the-art methods for text recognition in natural scene images. There are several methods [27-40] addressing different challenges posed by text in natural scene images. However, none of those state-of-the-art methods considers the complexity of tattoo text and hence, the scope of those methods is limited to regular scene text recognition.

B. End-to-end Approaches for Text Spotting in Natural Scenes

Liao et al. [41] developed an effective text spotting model in natural scene images based on a Segmentation Proposal Network. Wang et al. [42] proposed a model for addressing the challenges of text spotting in natural scene images. The approach effectively detects boundary points to identify and fit bounding boxes for text lines in any orientation. Qiao et al. [43] proposed a "text perceptron" end-to-end approach for arbitrarily shaped text spotting. Liao et al. [44] described a model for arbitrarily shaped text spotting in natural scene images. Liu et al. [45] explored an adaptive Bezier curve network for real-time end-to-end text spotting in natural scenes. Wang et al. [46] used a method based on kernel representation for accurate end-to-end text spotting in natural scene text images. Huang et al. [4] introduced the SwinTextSpotter framework, based on a transformer that unifies the text detection and recognition tasks for spotting. Zhang et al. [47]

introduced the Text Spotting Transformers (TESTR) framework. It employs a single encoder and dual decoders for text box control point regression and character recognition. Kittenplon et al. [48] proposed the TextTransSpotter, a text spotting approach based on a multi-task transformer employing weakly-supervised learning for text spotting. Ye et al. [49] also proposed end-to-end text spotting in natural scene images. Their approach uses a single decoder with explicit ordered points for text detection and recognition. The encoded points comprise text semantics and locations.

Although the above approaches employed transformers for text spotting in scene images, the methods are not effective for tattoo text spotting because of the challenges posed by the calligraphic style of tattoo text. In addition, the scope of the methods is confined to scene images and not tattoo images.

Furthermore, it is noted from the literature that GANs have been used successfully for image synthesis and for transforming text to image and vice versa, but not for tattoo text spotting [50-54]. Similarly, the Hilbert transform (HT) has been used for image recognition, fault detection, defect detection, and cognitive task understanding [8, 55-57]. Wang et al. [8] explored the HT for extracting general features (edges) in images but did not specifically focus on the problems of (tattoo or scene) text images. It is unclear for which specific applications those general image features may be most useful. The literature on the Hilbert transform shows that it has not been used for tattoo and scene text recognition yet.

In summary, it can be concluded that none of the state-of-the-art text spotting methods considers tattoo text images for spotting. Moreover, while Chowdhury et al. [58] very recently proposed a deformable convolutional and inception-based neural network (DCINN) for tattoo text detection, the scope of that work is confined to detection and does not extend to spotting. In addition, according to the literature the combination of Hilbert Transform and GAN has not been explored for either text spotting or tattoo text spotting.

III. Proposed Tattoo Text Spotting

The authors' objective has been to develop an end-to-end approach for text spotting in tattoos and scene images. The rationale is that an end-to-end spotting approach can produce accurate and reliable results by minimizing the adverse impact of errors in the intermediate steps of cropping, feature re-calculation, word separation, character grouping, and character segmentation [46].

The key challenges of tattoo text spotting are background complexity and the calligraphic style of the text. Moreover, the background design and the tattoo text may share the same color and texture properties. To deal with the challenges of identifying the fine details in the complex input images, the proposed approach first performs a Hilbert transform (HT), which enables the enhancement and retention of only the fine details in the images through the Optimum Phase Congruency (OPC) [9]. The rationale for proposing the use of the HT with OPC is as follows. The HT involves a fast Fourier transform, which is a well-known high-pass filter and hence the HT helps in enhancing the fine details in an image and suppressing the rest of the image. In the case of text, the phase information obtained by the HT results in high energy for edge pixels compared to the background, because high energy is represented by high frequency and amplitude. Therefore, the phase congruency exploits the high energy information to enhance and retain edge pixels by suppressing background pixels. This results in the fine details (edges) representing text information, irrespective of whether it is tattoo text or general scene text. In addition, the resulting edge detection in effect reduces the background complexity, regardless of the input image type (tattoo or general scene images).

Since the complexity of text spotting is reduced, the need for a large number of samples to train the model may also be reduced. This motivated the authors to introduce a GAN for text spotting in this work to further minimize the method's dependency on a large number of training samples. The GAN enables the generation of the necessary training text samples with fewer original samples (and not many are available in the case of tattoo text). In addition, the GAN can capture the geometrical structure and shape of characters in complex situations [9, 10]. In summary, the HT reduces background complexity, while the GAN simplifies dealing with the widely varying foreground complexity (calligraphic text style of tattoo text). Hence the proposed combination of HT and GAN achieves a generalization ability that results in higher performance in difficult situations. The generator of the GAN is primarily utilized for text detection and localization by generating synthetic images with various text patterns, styles, and deformations. These images aid in training the model to effectively identify and locate text regions within an image. The discriminator, on the other hand, is used for text recognition, ensuring that the generated images are appropriately recognized and transcribed into text (see Fig.2).

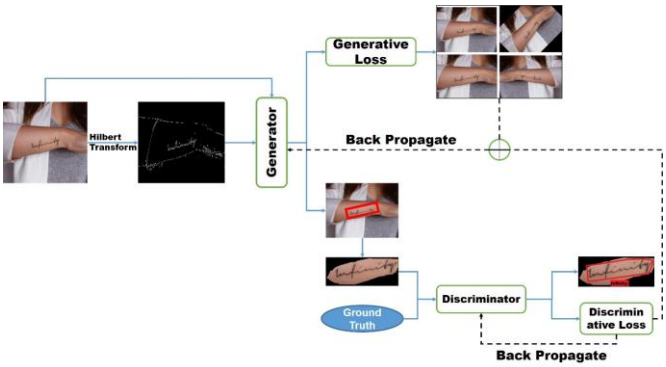


Fig. 2. Proposed text spotting framework for tattoo and natural scene images.

A. Hilbert Transform for Detection of Fine Details

As discussed in the previous section, background complexity is one of the key challenges for both tattoo and scene text spotting, and especially for arbitrary oriented and shaped text spotting. Motivated by the ability of the HT to suppress low frequency coefficient values (representing non-text pixels) and to retain high frequency coefficient values (representing text) irrespective of tattoo or natural scene images, the HT has been

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3378458

used for detecting fine details (edges) in the images. It is the authors' view that this approach is better than learning-based methods because the HT performs high pass filtering in an unsupervised way and hence it is a generalized approach for detecting the fine details in images of any type. The output of the HT also enables the subsequent steps of text detection and recognition (spotting) to perform well in both tattoo and natural scene images.

To achieve the above goal, the 2D Discrete Fourier Hilbert Transform (2D-DFHT) is computed to obtain the OPC for the input images.

To implement the 2D-DFHT in this work, the Riesz [59] transform is combined with the Discrete Fourier Transform (DFT). The objective functions of the Riesz and the DFT are defined in Eq.(1) and Eq.(2) respectively.

$$\left|\left|R_\mu\right|\right|^2_{L^2_\mu} = \iint_0^\infty \beta_{2,\mu}(p_x,r)^2 \cdot \frac{\mu(\beta(p_x,r))}{r^n} \cdot \frac{dr}{r} \cdot d\mu(p_x) \quad (1)$$

Here, $\beta(p_x,r)$ is the infimum taken over all neighboring n pixels whereas, $\beta_{2,\mu}(p_x,r)$ denotes its conjugate with mean distribution μ. Similarly, the DFT is defined as in Eq.(2):

$$B_j = \sum_{u=0}^{U-1} e^{-i\frac{2\pi ju}{U}} b_u \quad (2)$$

where, b_u depicts the cross-correlation between pixels with a complex sinusoidal function $e^{-i\frac{2\pi ju}{U}}$ over the pixel set U. Based on the above discussion, the pixel matrix $(N \times N)$ is analyzed in the spatial (odd pixel $(p_o(i,j))$) and in the frequency domain (even pixel $(p_e(i,j))$). Combining these two results in the 2D-DFHT as defined in Eq.(3).

$$\frac{p_o(i,j)}{p_e(i,j)} = \frac{|dsf(i,j)|}{B_j} + \left|\left|R_\mu\right|\right|^2_{L^2_\mu} \cdot abdy(i,j) \quad (3)$$

where, *dsf* refers to the 2D finite discrete signum function and the $abdy$ refers to the adjacent boundary pixel value defined in Eq.(4) and Eq.(5), respectively.

$$dsf(i,j) = \begin{cases} 1, & 0 < i < \frac{N}{2}, 0 < j < \frac{N}{2} \\ -1, & \frac{N}{2} < i < N, \frac{N}{2} < j < N \\ 0 & elsewhere \end{cases} \quad (4)$$

$$abdy(i,j) = \begin{cases} 1, & j = 0, 0 < i < \frac{N}{2} \\ -1, & j = 0, \frac{N}{2} < i < N \\ 1, & i = 0, 0 < j < \frac{N}{2} \\ -1, & i = 0, \frac{N}{2} < j < N \\ 0, & elsewhere \end{cases} \quad (5)$$

With these two pieces of information, the frequency spectrum is defined as

$$F(i,j) = [dsf(i,j) + abdy(i,j)] \cdot [p_o(i,j) + p_e(i,j)] \quad (6)$$

The spatial domain information is retrieved using Eq.(7):

$$S(i,j) = [p_o(i,j) + p_e(i,j)] \cdot \cot\left(\frac{\pi}{N}\right) p_o(i,j) + \tan\left(\frac{\pi}{N}\right) p_e(i,j) \quad (7)$$

With the above derivations, the input image of $(N \times N)$ is expanded as defined in Eq.(8):

$$P(i,j) = \frac{1}{N^2} \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} |F(u,v)| \cdot \sin(\varphi_{u,v}(i,j)) \quad (8)$$

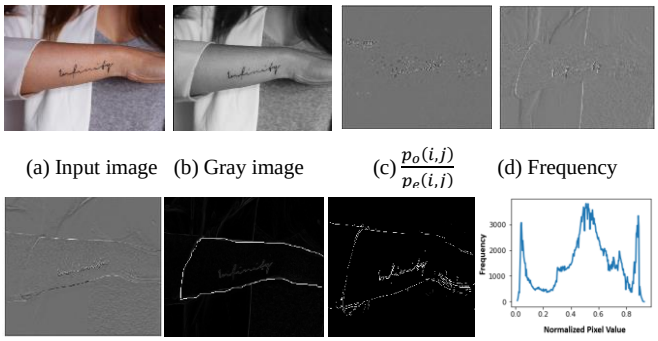
where $P(i,j) = p_o(i,j) + p_e(i,j)$ and $F(u, v)$ is the frequency spectrum distribution, and $\sin(\varphi_{u,v}(i,j))$ is the sinusoidal phase congruency between the i[th] and j[th] pixels. The local region of interest can be defined as in Eq.(9):

$$ROI(i,j) = \sqrt{P(i,j)^2 + \left(\frac{1}{N^2} \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} |S(u,v)| \cdot \cos\left(\varphi_{u,v}(i,j)\right)\right)^2} \quad (9)$$

Finally, the phase congruency information is extracted using Eq. (10):

$$\vartheta(i,j) = \frac{ROI(i,j)}{\frac{1}{N^2} \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} |F(u,v)| \cdot |S(u,v)|} \sin\left(\varphi_{u,v}(i,j)\right) \cdot \cos\left(\varphi_{u,v}(i,j)\right) \quad (10)$$

This phase information is convolved with the *Mf* and *Mh* [8] operators to produce the 0[th]-pixel phase information ($f_0(x,y)$) and the 1[st]-pixel phase information ($f_1(x,y)$) matrix, respectively. These are squared and summed up to obtain the local information energy. This information is divided by the obtained $ROI(i,j)$ to get the current candidate region.



(a) Input image (b) Gray image (c) $\frac{p_o(i,j)}{p_e(i,j)}$ (d) Frequency

(e) Spatial (f) Skin Region (g) Fine details (h) Histogram

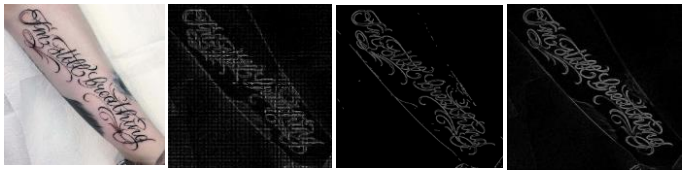
Fig. 3. Illustrating 2D-DFHT for obtaining candidate region.

The steps to obtain the OPC using the 2D-DFHT are illustrated in Fig. 3, where one can see (a) the input image and (b) the corresponding grayscale image. The effect of phase congruency can be seen in Fig. 3(c) and fine details in the frequency and the spatial domains can be seen in Fig. 3(d) and Fig. 3(e), respectively. Fig. 3(f) shows the skin region information in binary form. The final effect of the OPC can be seen in Fig. 3(g) as the fine details (edges).

The benefits of the Hilbert Transform step of the proposed method are evident from the histogram distribution of the OPC shown in Fig. 3(h), where the left and the right peaks represent the skin region boundary (edge) pixels while the highest peak represents edges of text. This example indicates the potential ability of the OPC to distinguish text from non-text information. The above observation is another justification for the use of HT and OPC for improving the performance of text spotting in tattoo and scene images.

When the HT is employed on the input image containing text, it generates a frequency coefficient matrix, where high and low frequency coefficient values can be seen. Since the HT behaves like a high pass filter, it discards low frequency coefficients which usually represent non-text information and retains high frequency coefficients which represent edge information as shown in Fig. 3(h). With this observation, the proposed method chooses high frequency coefficients which contribute to highest peaks in the histogram in the continuous domain, and the same coefficients are used to perform an inverse HT to detect the fine details (edges) in the spatial domain. Since the inverse transform chooses the coefficients which contribute to the majority for finding edge pixels, the gap between the continuous domain and the discrete domain does not affect edge detection. In this way, the proposed method facilitates domain transfer from the continuous frequency domain to the discrete spatial domain.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3378458



Input image Hough transform Canny Hilbert transform

Fig. 4. Comparing the proposed edge detection by Hilbert transform with other well-known edge detection methods.

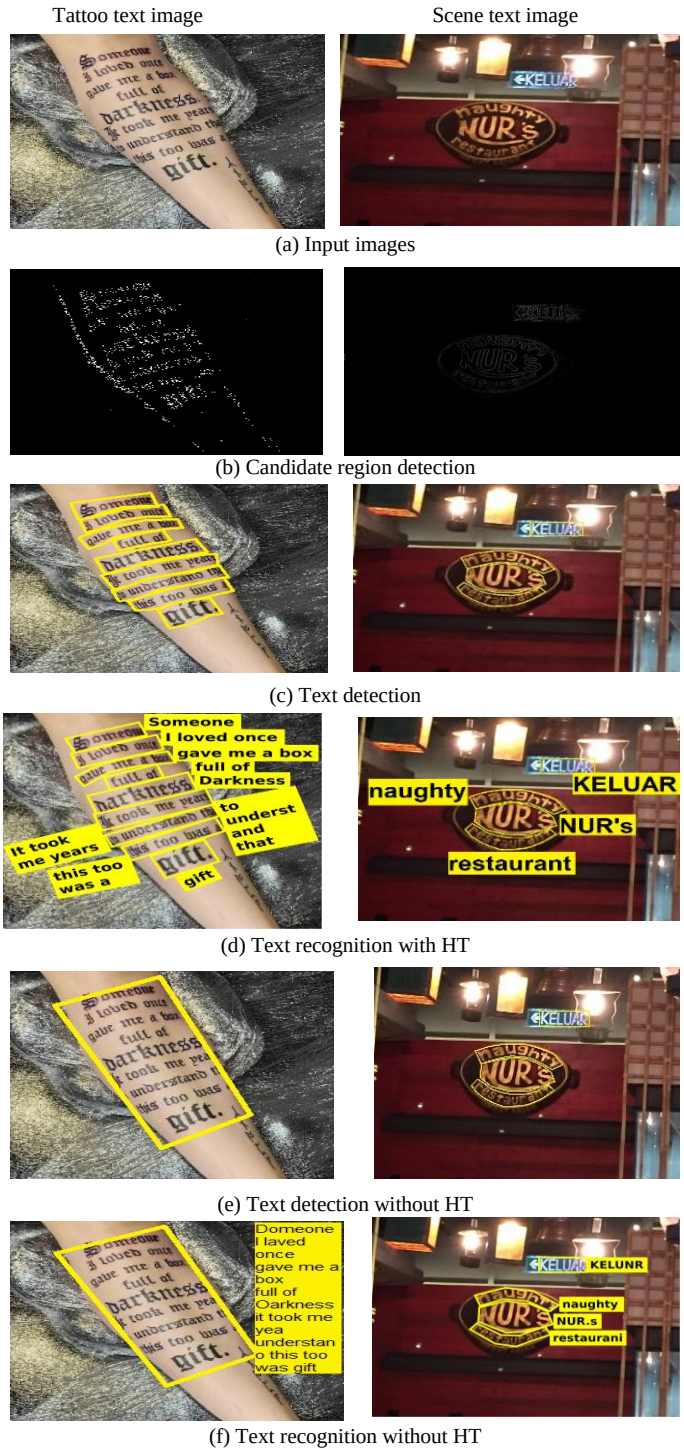
There are several well-known edge detectors in the literature, such as the Hough transform and Canny, which also provide fine details in the images. However, these are not suitable for tattoo images because of the unpredictable background and calligraphic style of tattoo text. It can be seen from the illustration shown in Fig. 4 for the input image, where Hough transform introduces noise, while Canny lost edge connectivity. The Hilbert transform (HT) enhances edges without introducing noise and loss of connectivity. The reason that the HT performs well is that very small and/or low contrast edges can be noticed in the frequency domain, in contrast to the spatial domain. Therefore, the HT is more suitable for the proposed work compared to the other techniques mentioned.

It is noted from Fig. 1 that the text in tattoo images can be difficult to distinguish from the background. In addition, since the tattoo text is similar to handwritten text – see also Fig. 5(a) – and the background is unpredictable, the overall challenge becomes greater. Moreover, the presence of tattoo text on different skin colors, different parts of the human body with different artistic backgrounds makes text spotting in tattoo images more complex compared to spotting regular text in natural scenes.

The effectiveness of the proposed HT for text spotting is illustrated in Fig. 5(a)-(f), where one can see for complex images of tattoo and scene text, the HT enables the reduction of background information by extracting the fine details which contain text information, as shown in Fig. 5(b). Similarly, the effectiveness of the fine detail extraction and background reduction achieved can be seen in Fig. 5(c) and Fig. 5(d), where the proposed method detects and recognizes all the tattoo text and the scene text accurately. In contrast, Fig. 5(e) and Fig. 5(f) show that the proposed method without the HT step does not detect and recognize the tattoo and scene text accurately. In the case of tattoo text, since bounding boxes are not identified accurately, the model fails to recognize all the text correctly. For the scene text, although bounding boxes are fitted adequately to the text, the model fails to recognize the word "KELUAR", recognizing it as "KELUNR" instead. Therefore, the above strongly indicates that the HT step is beneficial in addressing the challenges of both tattoo and scene text.

B. Tattoo Text Spotting

In this paper, the problem of text spotting is posed as an image-to-text translation task. The model needs to learn the mapping of the text boxes in the generator, given an input image. The discriminator uses those text boxes to recognize the text. The GAN architecture used in this work is shown in Fig. 2.



(a) Input images

(b) Candidate region detection

(c) Text detection

(d) Text recognition with HT

(e) Text detection without HT

(f) Text recognition without HT

Fig. 5. Illustrating the effectiveness of the HT for text spotting. Recognition results are shown in the yellow blocks.

The model used here is inspired from the pix2pix model for paired image-to-image translation [60]. The pix2pix model is based on cGAN [61]. Theoretically, as mentioned in that paper, translation is stated to be between two domains of images if they maintain a similar structure. The same concept has been explored in the method proposed here to perform image-to-text translation. The L1 loss along with the normal GAN losses are used in the pix2pix model.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3378458

The L1 loss prevents the GAN from producing completely new results, as the output textboxes must be tightly coupled, while the GAN loss accounts for accurate, non-blurry image-to-text spotting of the image. Since this work focuses on text spotting, in the proposed GAN architecture the image features are encoded and then decoded into another domain (text domain). To tackle different structures of cross-modalities, we introduce cGAN loss to ensure that the text aligns with the image (or vice versa). This can include alignment loss (to ensure image-text pairs match correctly), content loss (to ensure semantic content is consistent across modalities), and traditional adversarial loss [62]. More details for text spotting are presented below.

For each candidate region produced by the HT step, the *generator* of the GAN is used for detecting and localizing text, while the *discriminator* of the GAN is used for recognizing text. The target of the adversarial learning is to generate an accurate segmentation map, which is required for the subsequent stages: bounding box generation and cropping of the bounding box content. One of the novelties in this detection and recognition step is training the GAN by backpropagating a detection loss. In addition, the generator includes a double stacked hourglass network (SHG) [63] which generates a segmentation map, a center key point heatmap, and performs bounding box regression for accurate text detection, as shown in Fig. 6. Similarly, the discriminator includes a dynamic head attention network to recognize the text aided by ground truth information.

In Fig. 6, the SHGs are GAN-like subnets. Three subnet pairs are used: the first pair processes temporal and spatial information of the image, the second pair is used to process the output of the Hilbert Transform, contributing to the Image-to-Text domain adaptation. Last but not least, the third pair is used to combine all the features in order to generate bounding boxes.

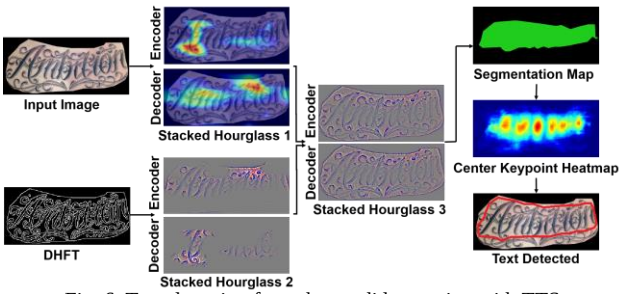


Fig. 6. Text detection from the candidate region with TTS.

a. Overview of the Training Schema

The proposed method trains a mapping from the domain input image (I) to the OPC information (ϑ). As mentioned earlier, to achieve this, the GAN-based architecture is employed as it can minimize the loss in bounding box generation and reconstruction in generators with multi-task learning [64] and a self-attention mechanism [65]. Since a pair of images are considered as the input, this mapping can be implemented in a fully supervised manner, with a reconstruction loss (ζ_r) as defined in Eq.(11):

$$\zeta_r = \sum_{i=1}^{N}\sum_{j=1}^{N} ||I(i,j) - \vartheta(i,j)|| \quad (11)$$

Furthermore, to prevent the generator from predicting overlapping bounding boxes, a perceptual loss (ζ_p) is adopted as defined in Eq.(12):

$$\zeta_p = \frac{1}{N^2}||\varphi(I(i,j)) \times \varphi(\vartheta(i,j))|| \quad (12)$$

where $\varphi(\cdot)$ evaluates the similarity between pixel values (resulting in 1 if similar, 0 if not). The similarity matrix of the input image is multiplied by the similarity matrix of the OPC information; hence, non-text pixels are assigned low values compared to text pixels. Non-text pixels can then be eliminated.

Additionally, an adversarial loss (ζ_A) [66] is utilized for fine-grained bounding box fitting improvement and ensures proper domain transfer from detection to recognition. It is worth mentioning here that the adopted discriminator function is based on both a local and a global discriminator. The global discriminator promotes better translation of the target domain (i.e., text recognition), while the local discriminator works on small fragments to ensure data retention (i.e., character recognition). To train the discriminator, the character-wise loss (ζ_D) is optimized, as defined in Eq.(13):

$$\zeta_D = \sum_t \ln P(y_t|I_t) \quad (13)$$

where, y_t and I_t represent the ground truth and the input of the t^{th} character, respectively. The loss function used in this work is discussed in detail in Section III.C.

b. Generator (G)

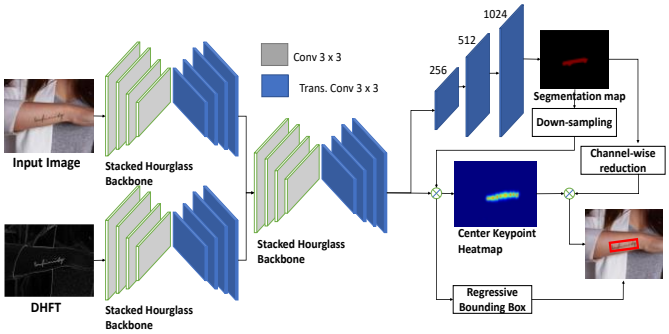
The illustration in Fig. 7(a) presents an overview of the generator architecture and its working principles. The input image and phase information (from the HT step) are fed into two parallel SHG networks for feature reconstruction. The output of these networks is supplied to the third SHG to perform the supervised mapping. The features are also passed through an up-sampling block consisting of three convolution layers with a kernel size of 3×3. The up-sampling layers are basically convolution layers with increasing numbers of kernels (i.e. 256, 512, 1024, and so on). It should be noted that for this overview, only the basic blocks are mentioned by which the up-sampling representation is made, referred to as a convolutional blocks. The outcome is the segmentation map for the input image. The results are then down-sampled and concatenated with the output of the third SHG to obtain the heatmap and regressive bounding boxes. Finally, these pieces of information are merged for detecting/localizing the text.

The down-sampling and the up-sampling have a number of advantages and some disadvantages, as discussed below. In case of down-sampling-**Advantages**: Down-sampling reduces the spatial resolution of feature maps but increases the receptive field of each neuron. This allows the network to capture larger contextual information, which can be beneficial for recognizing global patterns and objects. As the spatial dimensions decrease, the number of parameters and computations in subsequent layers also decreases. This reduction in complexity can lead to faster training and inference times. Down-sampling is often used in a hierarchical manner, where lower layers capture fine-grained details, and higher layers capture more abstract and global information as shown in Fig. 7(b). This hierarchical feature learning can be advantageous for object detection. **Disadvantages**: Down-sampling discards fine-grained spatial information, which can be crucial for tasks that require precise localization of objects. In object detection, this loss of

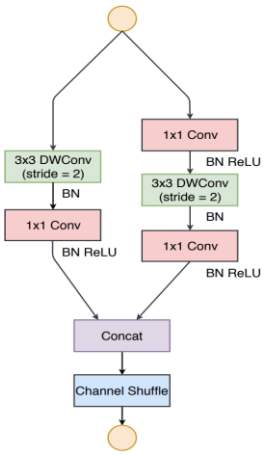
This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3378458

information may affect the accuracy of bounding box predictions.

In the case of up-sampling–**Advantages**: Up-sampling helps in recovering the spatial information lost during down-sampling. This is crucial for tasks like object detection where precise localization is essential. Up-sampling allows the network to generate high-resolution feature maps, improving the precision of object localization. This is particularly important for detecting small objects. **Disadvantages**: Up-sampling often involves introducing more parameters and computations, leading to increased computational complexity. This can result in longer training and inference times. The introduction of more parameters during up-sampling may increase the risk of overfitting, especially if the dataset is limited.



(a) The generator architecture.



(b) The down-sampling block.

Fig. 7. Illustration of the generator architecture.

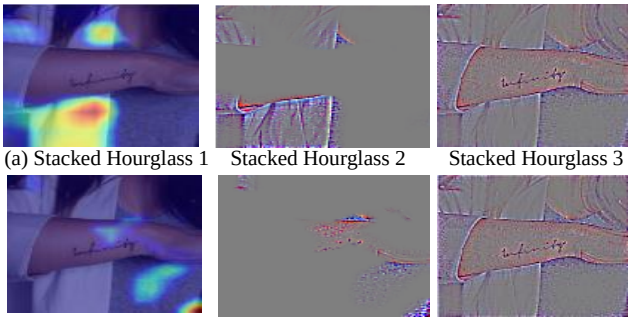
Baseline Architecture: This approach trains the SHG [63] to identify the key points of characters, such as intersection and junction points, by aligning (by minimizing the loss – distance between the center points) the center of the regressive bounding box to the center of the heatmap obtained from the phase congruency. The height and width of the bounding box are regressed as well as the offset between the boxes. The result is, therefore, the maximum area heatmap for each possible label, the region size for each point, and the offset for each point, as shown in Fig. 7.

Multi-Task Learning (MTL): The network is trained to create the segmentation of a text region while discovering its bounding box, thus it is important to make shared parameters more comprehensive and to prevent overfitting. To do this, a two-phase partition (encoder stack/decoder stack) has been inserted into the network. This architecture can be likened to a sequence-to-sequence model commonly used in natural language processing tasks such as text generation. The encoder and decoder stacks are trained using a semi-supervised learning approach where generated text tokens are used as labels during training [67]. The model aims to generate text or translations that align with the provided tokens, incorporating these descriptors into the generated output. The additional subdivision head considers the inclusion of a feature map that has been reduced by a quadruple spatial location process (splitting the information into four equal pixel matrices) compared to the input. It has 3×3 convolutions, with high layers in the middle. The channel size is reduced to 1 in the final agreement, thus leading to a split map with the same width and height as input, per channel. Here, the adversarial loss (ζ_A) is used, as defined in Eq. (14):

$$\zeta_A = \frac{1}{N}\sum_{i=1}^{N} \beta_i(\mathrm{p_i}\log(p) + p_i^*\log(1-p)) \qquad (14)$$

where $\beta_i = \frac{N}{G(p_i)^\gamma}$, p_i is the pixel value, G is the generative function, and γ is the normalized mean pixel value.

Self-Attention Technique (SAT): It receives as input the segmentation map, down-sampled by the generator by a rate of 4 to reduce it to the location limits of the first element map. To minimize feedback on areas that may contain useful information, this technique replicates the entire feature map channel with the obtained split map, thus reducing the likelihood of false detections in unrelated locations.



(a) Stacked Hourglass 1 Stacked Hourglass 2 Stacked Hourglass 3

(b) Stacked Hourglass 1 Stacked Hourglass 2 Stacked Hourglass 3

Fig. 8. Activation map of the SHGs. (a) Encoder and (b) Decoder.

Semi-Supervised Annotations (SSA): To make the generator more robust for automated text detection and to ensure the accuracy of bounding boxes, the results of the 2D-DHFT are intersected with the obtained segmentation map to achieve noise reduction as well as pixel-wise heatmap annotation. This enables the generation of tight-fitting arbitrarily shaped bounding boxes.

The activation map results from the three SHGs used in the generator are illustrated in Fig. 8. It can be seen that incomplete segmentation annotations not only are good enough to train useful attention maps, but they also allow the regressive head to produce an accurate bounding box compared to the state-of-the-art (e.g., see Fig. 1).

It should be noted that in the first SHG (where the input

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3378458

images are fed), neither encoder nor decoder activates the ROI, while the decoder of the second SHG (where the phase information has been fed) attempts to generate an ROI, albeit not accurate. Subsequently, when these two pieces of information are passed to the third SHG, the results of the encoder as well as of the decoder accurately generate the ROI. This is the key to the effectiveness of the combination of multi-task learning, self-attention, and self-supervised annotation in the generator.

c. Discriminator (D)

The output feature tensor from the generator is considered along with the ground truth information in order to validate it. The detailed architecture of the discriminator network can be seen in Fig. 9. In this discriminator, the input tensor ($\tau \in \mathbb{R}^{R\times G\times B}$) is extracted from the detected text region to obtain the objective function of the discriminator (D), as defined in Eq. (15):

$$D(\tau) = \delta(\tau) \cdot \tau \quad (15)$$

where $\delta(\tau)$ is the cumulative attention function. This circumvents the use of fully connected layers, which also helps to reduce the time complexity. To achieve this, $\delta(\tau)$ is expanded as defined in Eq. (16):

$$\delta(\tau) = \delta_T(\delta_{SP}(\delta_{SC}(\tau)\cdot\tau)\cdot\tau) \quad (16)$$

where δ_T, δ_{SP}, and δ_{SC} denote task-aware, spatial-aware, and scale-aware attention, respectively as discussed in detail below.

Scale-aware Attention (ScAT): This attention function measures dynamic elements in different proportions according to their semantic significance as detailed in Eq. (17):

$$\delta_{SC}(\tau) = \sigma(L(\frac{1}{N}\sum_{i=1}^{N}\tau * (1-\tau))) \quad (17)$$

where $L(\cdot)$ is the linear activation function incorporated with a 1×1 convolutional layer and $\sigma(y) = \max(0, \min(1, \frac{y+1}{2}))$ provides a measurement of feature similarity to follow the text curvature.

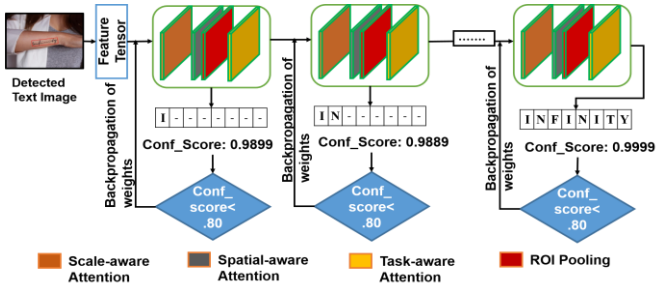


Fig. 9. The architecture of the discriminator block.

Spatial-aware Attention (SpAT): To strengthen the features for recognition, an alternative location-focused module has been introduced which is based on integrating the identified characters into words/strings by concentrating the text region fragments (characters) between naturally delimiting regions (space). By assessing the maximum size of the feature tensor after scaling, this module decomposes the feature tensor (output by the generator) into smaller components of spatial information into two steps: first makes attention accessible through a deformable convolution and then integrates features at all levels into the same spatial tensor, as depicted in Eq.(18):

$$\delta_{SP}(\tau) = \frac{1}{len}\sum_{i=1}^{len}\sum_{j=1}^{SSL} w_{i,j}\cdot\left((p_{SSL}+\Delta p_{SSL})\right)^{i}\cdot(\nabla m_{SSL})^{j} \quad (18)$$

where len represents the length of the feature vector after scaling, SSL denotes sparse sampling locations, $p_{SSL}+\Delta p_{SSL}$ represents the deviation from the recognized region rectified by the self-learned spatial offset, and ∇m_{SSL} represents the loss of self-learning in feature recognition. After that, an ROI integration layer has been implemented to extract intermediate representations from the previous layers and maintain them in the spatial tensor.

Task-aware Attention (TAT): To be able to learn collaboratively and practice a common representation, this attention function has been utilized to recognize text in the end. It drastically changes the recognized characters' features to allow grouping functions for word formulation into a single line. The objective function of the task-aware attention is given in Eq. (19):

$$\delta_T(\tau) = \max\left(\alpha\delta_{SP}(\tau)+\beta^2\delta_{SP}(\tau), \alpha^2\delta_{SP}(\tau)+\beta\delta_{SP}(\tau)\right) \quad (19)$$

where, α, β are hyperparameters, initially starting with (-1,-1), chosen according to the best experimental scenario as $\delta_{SP}(\tau)$ is implemented with self-learning. To implement the function, a global average pooling combination of size $len \times SSL$ is first performed to reduce the size of the tensor. Then the resulting tensor is passed through two fully connected layers followed by a batch normalization and ultimately activates the modified sigmoid function – the hyperbolic tangent (tanh) function used to normalize the results within (-1, 1). An overview of how this three-attention network works simultaneously is shown in Fig. 10.

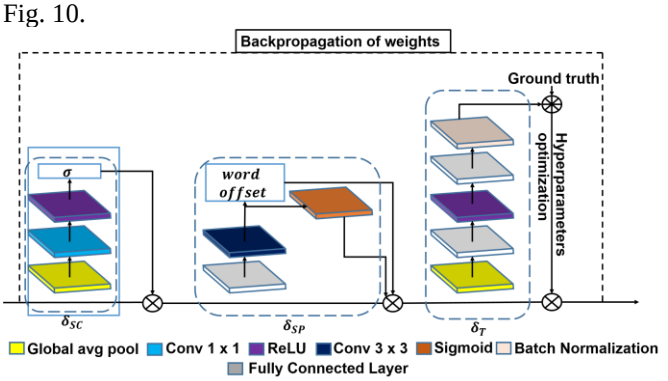


Fig. 10. Working principle of the δ_T, δ_{SP}, and δ_{SC} combination.

Fig. 11. Step-by-step character recognition by the proposed model.

The recognition network takes the region features of the cropped images as input. The bounding boxes obtained by the generator are cropped through adversarial mapping. Now the loss function of the discriminator is the character error rate (CER). If the CER increases, then ζ_D, defined in Eq. (13), also

increases. Similarly, the adversarial loss (ζ_{GAN}) also increases. In the network, this adversarial loss works as a gradient and flows to the generator to change the weights of the generator. As ζ_{GAN} is the single loss function of the overall network, the cropping is bypassed and the training is end-to-end.

The step-by-step text recognition process is illustrated in Fig. 11, where the input is a text region with a bounding box, and recognition is performed character by character. It should be noted that a significant advantage of the proposed model is that characters are accurately recognized even when there is no space between them.

C. cGAN Generator Loss

To better address the challenges of tattoo and scene text, a new loss function is proposed here, namely the *cGAN generator loss*. The objective function is defined similarly to cGAN which can be expressed as in Eq. (20), where G tries to minimize the objective against an adversarial D that tries to maximize it.

$$L_{cGAN}(G,D) = E_{x,y}[\log D(x,y)] + E_{x,z}[\log(1 - D(x,G(x,z))] \quad (20)$$

where $G^* = arg\min_G \max_D L_{cGAN}(G,D)$.

It is stated in the literature that it is beneficial if the GAN losses are combined with some traditional losses, such as the L1 loss [60]. In this work therefore, the L1 loss and the L2 loss are used as two additional losses apart from the cGAN losses. The L1-norm is also known as least absolute deviations (LAD) or least absolute errors (LAE). The L1-norm minimizes the sum of the absolute differences between the target values and the estimated values. The L2-norm is known as least squares, and it minimizes the sum of the squares of the differences between the target values and the estimated values. The task of the discriminator remains the same, whereas the task of the generator is not only to outwit the discriminator but also to stay near the ground truth in terms of L1 and L2. The mathematical representations for the L1 and L2 losses are shown in Eq. (21) and Eq. (22) respectively. The final objective function of G can be represented as shown in Eq. (23).

$$L_{L1}(G) = E_{x,y,z}[||y - G(x,z)||_1] \quad (21)$$

$$L_{L2}(G) = E_{x,y,z}[||y - G(x,z)||_2] \quad (22)$$

$$G* = arg\min_G \max_D [gan_{weight} * L_{cGAN}(G,D) + l_{weight} * (L_{L1}(G) + L_{L2}(G))] \quad (23)$$

where, gan_{weight} and l_{weight} are the corresponding ratios in which the GAN losses and the normalization losses are considered. The performance improvement achieved by using this new loss is demonstrated and discussed in the experimental results section (Section IV).

D. End-to-End Training

It is true that text detection performance improves with the help of recognition (eliminating false positives) and, at the same time, text recognition performance improves with the help of text detection (tight-fitting bounding boxes for the text lines). To achieve this, the model should be trained with text detection and recognition samples. Therefore, the proposed model shares the training samples to achieve end-to-end performance. More specifically, this is realized as follows.

End-to-end training refers to a complex learning paradigm, treating the whole as a single neural network (criterion 1) bypassing the intermediate layers (criterion 2). In the proposed model, the discriminator network loss function helps to improve the text detection (i.e., the loss will increase if the generator fails to generate accurate and tight bounding boxes); similarly, the generator network loss function helps to improve the recognition performance of the discriminator (i.e., if the bounding box is not accurate, it leads to the output of garbage characters as text recognition results). In the training phase of the proposed method, the discriminator and generator networks are trained sequentially to improve the performance of both the discriminator and the generator. Here both depend on each other to improve their performance through a single adversarial loss function that meets criterion 1 (single neural network paradigm), bypassing the cropping of the bounding boxes through the adversarial mapping, which meets criterion 2. It can therefore be seen that the training achieved is end-to-end.

IV. Experimental Results

To validate the proposed method in the case of tattoo text spotting, a new dataset has been created due to the lack of suitable standard datasets available in the literature for the specific challenges of tattoo text. This dataset has been made publicly available (details below). To validate the proposed method in the case of scene text spotting, the most prominent existing scene text benchmark datasets were used.

A. Datasets and Evaluation

Tattoo Text Spotting Dataset (TTSD): This newly created dataset comprises 500 RGB images for experimentation. Images are collected from different internet resources, such as social media as well as captured by the authors. This dataset includes representative images with several challenging characteristics as compared to natural scene images, such as freestyle writing, unusual character shapes (e.g., calligraphic characters), dense text lines with decorative letters, symbols, and decorative designs in the background. All the tattoo text is written in English, numbering 3524 text instances in total (having a mean of 5 text instances per image) with word-level annotation. The new dataset can be freely downloaded[1].

Scene Text Datasets: To demonstrate the effectiveness of the proposed model for detecting and spotting text in natural scene images in the presence of different challenges, such as arbitrarily oriented and shaped text, the text of multiple scripts, and curved text, the following benchmark datasets were used in experiments: CTW1500 [68], Total-Text [69], and ICDAR 2015 dataset [70]. The reason for choosing these datasets is that these are popular standard datasets for text spotting in literature. Furthermore, these datasets provide ground truth for both detection and recognition.

To evaluate text detection, the standard measures were used, Recall (R), Precision (P), and F-measure (F), and the same evaluation scheme was followed for calculating the metrics. In order to evaluate the end-to-end text spotting, the end-to-end (E2E) accuracy has been calculated. For TTSD, Total-Text and CTW1500, the E2E accuracy has been calculated in two

[1]https://drive.google.com/drive/folders/1o4WYa0gXuFWx6hlEGnpaZV-sEz8zElB4?usp=sharing

categories: *None* and *Full*. Here, *None* refers to recognizing the text without lexicon information, while *Full* refers to the lexicon which contains all the words in the test set. For ICDAR 2015, the E2E accuracy has been calculated in 3 categories regarding the lexicon: *Strong (S)*, *Weak (W)* and *Generic (G)*.

To show the effectiveness of the proposed method, we consider the state-of-the-art methods of text spotting in natural scene images for comparative study [1, 4, 14, 15, 45, 47, 48, 49]. The motivation to choose the above methods for comparative study with the proposed method is that the objective of text spotting in natural scenes is the same as the proposed method.

Training/Evaluation Details: The proposed method is developed with the ADAM optimizer and is trained and tested with several different training and testing samples of the respective datasets. Initially, data augmentation was performed during training by random resizing. Here, the shorter edges are kept in the range of 512 to 1024, while the most extended edges are kept in the range of 1024 to 2048. Besides that, instance-aware random cropping has also been performed, which ensures the cropped size is larger than half of the original size and no text regions being cut. However, the maximum image size is always fixed during testing at 1024.

With this configuration, the model was trained on the TTSD and Total-Text datasets for 100k epochs with a consistent learning rate of 10^{-5} and was decayed at the 40k-th iteration by a factor of 0.01. Not only that, but the learning rates were also scaled by a factor of 0.01 for the non-linear projections used to predict reference points as well as sampling offsets of the task-aware attention. The proposed end-to-end model has been optimized with the Adam optimizer as mentioned, with a weight decay of 10^{-3}. The training process takes about 3 days on 4 TITAN A40 GPUs with an image batch size of 64. The code is freely available[2].

For validation, 25% of the total dataset was used with a learning level and pre-configuration process for better performance. In addition, as the proposed network has a generator module that can increase the resolution in each epoch, the input resolution does not affect the end-to-end performance.

The same experimental setup was adapted for all state-of-the-art methods implemented for comparative study in this work. It should be noted those methods were trained from scratch and no pretrained model has been used for the experiments. However, we pretrained the proposed method on the TTSD dataset and fine-tuned on respective individual datasets for 20k epochs.

To demonstrate that the proposed model is generic, and that its performance does not depend on a large number of samples, the proposed model uses the samples from TTSD and Total-Text as well as the samples generated by different augmentation techniques for training. The key reason for not requiring many samples is that the HT step of detecting fine details helps in reducing the complexity of the problem and the proposed GAN helps in generating text samples (possible variants of tattoo text images) automatically.

All the state-of-the-art methods listed in Table 5 are fine-tuned with the samples of TTSD dataset before calculating the evaluation measures. Even though the same experimental set up is followed for both the proposed and the state-of-the-art methods for all the experiments, the state-of-the-art methods perform worse for TTSD compared to the proposed method. This may also be due to the small number of samples in the TTSD dataset affecting training. On the other hand, since the proposed method is effective for small and large numbers of samples, it achieves the best results for both the TTSD and other datasets.

B. Ablation Study

In this work, the stacked hourglass (SHG) networks are used to encode the image features in such a way which enables text spotting via image to text translation. As discussed earlier, three sets of SHG encoder-decoder pairs were used to encode the temporal and spatial information (via the first SHG pair), to perform image-to-text domain adaptation through the second SHG pair, and to reconstruct an image which contains more information than the raw input one (a super annotated image [71]) using the third SHG pair which combines all the features. This enables the avoidance of pre-training using very large datasets such as SynthText, without compromising accuracy. To demonstrate that the SHG architecture is effective and contributes to the proposed method's improved performance in text detection and spotting, experiments were conducted using different backbone architectures, comparing them with the proposed SHG+GAN on the TTSD dataset. The results are reported in Table 1, where it can be seen that the proposed SHG+GAN achieves the best text detection results (P, R and F) and the best spotting results (Full) compared to other backbone architectures. The other combinations do not provide satisfactory results because they do not have the positional embedding power of the proposed method.

Table 1: Experiments on different backbone architectures and the proposed SGH + GAN using TTSD.

Methods	P	R	F	None	Full
ResNet-50 + GAN	84.8	82.0	83.3	75.1	72.6
ResNet-101+ GAN	86.0	81.1	83.4	**86.3**	83.0
VGG-16+ GAN	85.9	80.1	83.2	69.2	74.4
DenseNet + GAN	85.9	83.4	84.6	78.2	81.0
SHG + GAN	**93.3**	**93.1**	**93.1**	85.7	**89.4**

As mentioned earlier, the proposed model uses a new loss function, the *cGAN generator loss*, which is a combination of the L1, the L2 and the GAN loss. To demonstrate that the proposed loss is better than L1, L2 and their combination, experiments were conducted on the TTSD dataset and the performance of the model using different losses is presented in Fig. 12. It is observed from Fig. 12 that the combination of cGAN + L1 + L2 results in the best performance compared to other individual losses and combinations. This shows that the proposed cGAN generator loss is effective. Using the other losses results in poorer performance because sometimes using

[2]https://drive.google.com/drive/folders/18WqBrhcEiIBAEMYNZiiWMs6FZOBhDBwJ?usp=sharing

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3378458

L1 ignores important features while the L2 loss magnifies the error if the model makes a single poor prediction. To overcome these drawbacks of the single regressive loss, the cGAN generator loss combination is used in this work.

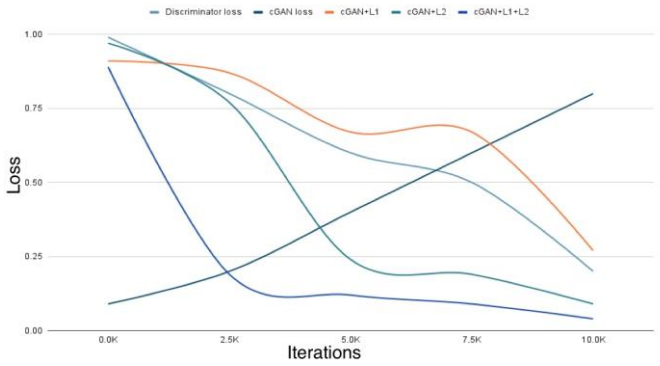


Fig. 12. The effectiveness of the proposed cGAN generator loss plotted against the individual losses and combinations.

The results of different ablation experiments, (i) to (xi), on the TTSD dataset and on a standard scene dataset (Total-Text) are reported in Table 2. Since the Total-Text dataset contains several images (1,555) with diverse text instances, it covers a wide range of text variations including different languages, orientations, shapes, sizes and backgrounds. Therefore, this dataset is well-suited for training and evaluating models in handling challenging text layouts and deformations often encountered in natural scenes.

It Is observed from the experiments (i) to (ix) in Table 2 that the results of text detection and text spotting performance improve gradually when the individual steps of the proposed method are added one by one, compared to the results of the baseline architecture – just the GAN in (i). This shows that all the proposed components make valid and effective contributions in improving the performance of the proposed method in both text detection and spotting.

Comparing (viii) and (xi), it can be observed that using the HT step produces better results than using the Fourier transform (FFT), for both text detection and spotting. This shows that the HT step is more effective for detecting the fine details than the FFT. Similarly, when the results of experiments (ix) and (x), where the proposed method skips Optimum Phase Congruency (OPC) step and uses RGB images instead of grayscale images, respectively are compared with the results of the proposed method (xi) for text detection and spotting, the proposed method achieves better results. Therefore, using the OPC and grayscale images are important considerations and effective for tackling the challenges of tattoo and scene text spotting. The reason for the poor results in the case of using RGB images compared to grayscale is that this introduces unnecessary complexity when in fact text does not usually exhibit large variations in color. For instance, the whole text line will have very similar color and texture. However, the orientation, scale and aspect ratio of each character varies.

Table 2: Ablation study of the proposed method using the TTSD and the Total-Text dataset.

Exp.	Key Steps	TTSD					Total-Text dataset				
		Text Detection			Text Spotting (End-to-End)		Text Detection			Text Spotting (End-to-End)	
		P	R	F	None	Full	P	R	F	None	Full
(i)	Base line: GAN	84.3	90.2	87.8	65.3	77.4	62.1	45.5	52.5	52.9	71.8
(ii)	GAN+δ_{SC}	84.4	90.1	87.1	70.4	78.1	69.0	55.0	61.3	55.8	79.2
(iii)	GAN+δ_{SC}+δ_{SP}	84.7	90.4	87.7	71.2	78.4	72.5	83.4	77.6	57.5	77.2
(iv)	GAN+δ_{SC}+δ_{SP}+δ_T	85.1	91.2	88.2	73.5	80.7	72.1	84.6	77.9	58.4	79.0
(v)	GAN+δ_{SC}+δ_{SP}+δ_T +MTL	87.4	91.8	89.6	74.9	83.6	85.6	75.7	80.3	65.0	76.1
(vi)	GAN+δ_{SC}+δ_{SP}+δ_T +MTL+SAT	89.7	91.9	90.5	76.3	83.9	88.8	81.8	85.2	69.7	78.3
(vii)	GAN+δ_{SC}+δ_{SP}+δ_T +MTL+SAT+SSA	91.2	92.7	91.9	77.6	86.2	88.9	85.0	87.0	72.9	83.6
(viii)	(Proposed without HT) + FFT	91.4	92.6	91.4	80.3	87.1	92.8	83.7	88.0	73.3	83.9
(ix)	Proposed w/o OPC	91.0	89.7	90.3	83.9	87.2	92.9	84.7	88.6	77.1	84.6
(x)	Proposed-RGB	86.9	89.1	87.9	81.2	84.3	92.1	80.2	85.7	76.7	81.2
(xi)	Proposed	**93.3**	**93.1**	**93.1**	**85.7**	**89.4**	**93.4**	**85.2**	**89.1**	**77.3**	**86.1**

To validate the general advantages of the proposed HT step itself, experiments were conducted by feeding the output of the HT step to DeepSolo [49], a very recent state-of-the-art method for text spotting in natural scene images. Evaluation measures were calculated on the TTSD and the Total-Text dataset. The results are reported in Table 3, where it is noted that the performance of DeepSolo with HT is better than the performance of DeepSolo without HT. Thus, we can conclude that the proposed HT for the detection of fine details is important for achieving higher performance and it is generic.

In addition, to show that the proposed model does not depend on a large number of training samples, especially on the SynthText dataset, experiments were conducted by training on SynthText first and then fine-tuning on respective datasets. The results reported in Table 4 show that the performance of the proposed method is better when *not* using SythText and Fine-tuning on all the four datasets. To probe further, the proposed method is also compared with the SADA-SSC state-of-the-art method in [13], which proposes multi-scale context aware feature aggregation for curved scene text detection and does not use SynthText either. The results in Table 4 indicate that the performance of the proposed method is superior either with or without SynthText and fine-tuning training. Therefore, one can assert that the proposed method is capable of achieving superior results for text spotting in tattoo and scene images without the use of a large number of samples.

Table. 3. Experiments using DeepSolo [49] with and without the proposed HT step on different datasets.

Dataset	DeepSolo with HT					DeepSolo without HT				
	Text Detection			Text Spotting		Text Detection			Text Spotting	
	P	R	F	None	Full	P	R	F	None	Full
TTSD	86.7	78.2	82.2	49.1	76.4	85.6	75.7	80.3	48.8	74.8
Total-Text	93.1	87.6	90.2	83.9	89.8	92.9	87.4	90.0	83.6	89.6

Table. 4. Experiments on different training strategies for text detection.

Training Strategy	TTSD			Total-Text			CTW1500			ICDAR15		
	P	R	F	P	R	F	P	R	F	P	R	F
Proposed with SynthText + Fine-tuning	84.6	89.2	86.8	91.1	86.2	88.5	86.4	81.2	83.7	88.6	87.5	88.1
SADA-SSC [13]	81.7	78.2	79.9	86.7	82.6	84.6	87.2	81.7	84.4	88.8	82.6	85.6
Proposed	**93.3**	**93.1**	**93.1**	**93.4**	**85.2**	**89.1**	**92.2**	**84.4**	**88.6**	**89.7**	**91.4**	**90.5**

C. *End-to-End Experiments for Text Spotting*

Qualitative results of the proposed and state-of-the-art methods for text spotting on the TTSD and the different benchmark natural scene datasets are shown in Fig. 13 and Fig. 14, respectively. It can be observed that the proposed method spots both tattoo and scene text accurately, indicating that it is independent of image type. On the other hand, although state-of-the-art methods *detect* tattoo and scene text well, they fail to

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3378458

accurately *spot* the text in both types of images. This shows that the state-of-the-art methods are limited – confined to scene text spotting and sensitive to low contrast and small font size.

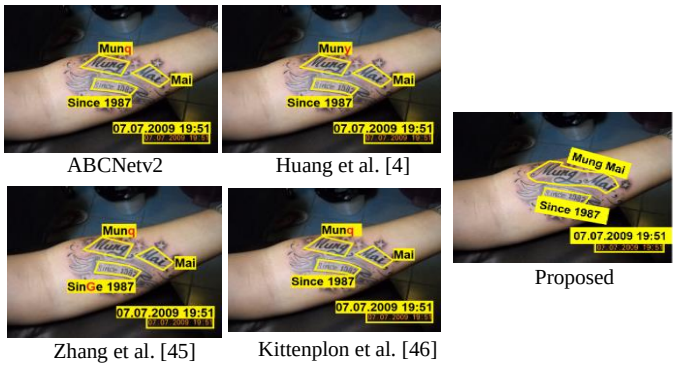


Fig. 13. Text spotting on TTSD dataset. Recognition errors shown in red.

Fig. 14. Text spotting on benchmark scene text datasets. Recognition errors shown in red.

Quantitative results of the proposed and the state-of-the-art methods on the TTSD and the three scene text benchmark datasets (Total-Text, CTW1500 and ICDAR 2015) are reported in Tables 5 to 8, respectively. For the TTSD, the proposed method performs best in both detection and spotting (End-to-End) compared to the state-of-the-art. The main reason for the poorer results of the state-of-the-art methods is that the scope of those methods is limited to regular text in natural scenes. While the state-of-the-art methods are effective for complex text and backgrounds and achieve reasonably good accuracy, it is not sufficient to beat the accuracy of the proposed method. In the case of the proposed system, the HT step for detecting candidate regions and the use of the GAN to achieve an end-to-end model prove to be an advantage.

When the performance of the proposed method is considered against the state-of-the-art on text detection and spotting in natural scene images, as reported in Tables 6 to 8, the proposed method does not always outperform but it is among the best overall. More specifically, for text detection, the proposed method achieves the best Precision on the Total-Text dataset, the best precision and F-measure on the CTW1500 dataset and the highest Recall and F-measure on the ICDAR 2015 dataset. In the case of text spotting in natural scenes, the proposed method is the second highest in terms of both *None* and *Full* on the Total-Text dataset, second highest in terms of *None* and third highest in terms of *Full* on the CTW1500 dataset, and the second highest in all cases (*Strong, Weak and Generic*) on the ICDAR 2015 dataset, compared to the state-of-the-art.

Since the primary focus of the proposed method is to achieve the best results in the new application area of tattoo text spotting, the slightly lower than the state-of-the-art performance on text detection and spotting in natural scene images is considered understandable and acceptable. Moreover, as the state-of-the-art methods are inferior to the proposed method for tattoo text detection and spotting and considering the overall performance of the proposed system on both tattoo text and natural scene text, one can reasonably conclude that the proposed system is generally superior.

Table 5: Comparative study on TTSD dataset for detection and spotting.

Methods	Text Detection			Text Spotting (End-to-End)		Response Time
	P	R	F	None	Full	
ABCNetv2 [45]	73.7	74.3	74.0	62.7	65.4	1.7
Huang et al. [4]	66.8	88.5	76.1	68.6	78.6	1.5
Zhang et al. [47]	92.9	77.8	84.7	67.5	73.3	1.8
Kittenplon et al. [48]	83.9	87.3	85.6	70.2	77.1	2.1
PAN++[1]	82.6	72.9	77.4	69.8	77.4	-
DBnet [22]	85.2	85.5	85.1	-	-	-
BRN-BCTS [14]	78.9	72.1	75.3	-	-	-
RCLM [15]	82.1	83.2	82.6	-	-	-
MaskTextSpotter v3 [41]	86.1	86.4	86.2	83.8	86.5	-
East [23]	32.9	33.3	33.1	-	-	-
SegLink [24]	83.8	84.2	84.0	-	-	-
MOST [25]	81.7	80.2	81.2	-	-	-
DeepSolo [49]	85.6	75.7	80.3	48.8	74.8	4.9
Proposed	**93.3**	**93.1**	**93.1**	**85.7**	**93.3**	**1.3**

Table 6: Comparative study on Total-Text dataset for detection and spotting

Methods	Text Detection			Text Spotting (End-to-End)		Response Time
	P	R	F	None	Full	
ABCNetv2 [45]	84.1	**90.2**	87.0	73.5	80.7	1.5
Huang et al. [4]	87.5	88.5	88.0	74.3	84.1	3.8
Zhang et al. [47]	92.0	82.6	87.1	73.3	83.9	2.2
Kittenplon et al. [48]	88.4	82.8	85.5	75.6	84.4	1.8
PAN++[1]	89.9	81.0	85.3	68.6	78.6	-
DBnet [22]	88.9	83.2	86.0	-	-	-
BRN-BCTS [14]	78.8	70.6	74.5	-	-	-
RCLM [15]	88.5	82.0	85.2	-	-	-
MaskTextSpotter v3 [41]	-	-	-	71.2	78.4	-

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3378458

East [23]	43.2	27.1	33.3	-	-	-
SegLink [24]	86.2	70.2	77.1	-	-	-
MOST [25]	90.4	82.7	86.4	-	-	-
DeepSolo [49]	92.9	87.4	**90.0**	**83.6**	**89.6**	2.2
Proposed	**93.4**	85.2	89.1	77.3	86.1	**1.1**

Table 7: Comparative study on CTW1500 dataset for detection and spotting

Methods	Text Detection			Text Spotting (End-to-End)		Response Time
	P	R	F	None	Full	
ABCNetv2 [45]	83.8	85.6	84.7	58.4	79.0	2.6
Huang et al. [4]	86.8	**89.2**	88.0	51.8	77.0	**1.6**
Zhang et al. [47]	92.0	82.6	87.1	56.0	**81.5**	2.7
Kittenplon et al. [48]	91.4	85.1	88.1	55.4	71.9	2.2
PAN++[1]	87.1	81.1	84.0	54.9	75.7	2.6
DBnet [22]	87.9	82.8	85.3	-	-	-
BRN-BCTS [14]	72.3	69.4	70.8	-	-	-
RCLM [15]	86.1	81.2	83.7	-	-	-
East [23]	0.25	0.25	0.25	-	-	-
SegLink [24]	87.7	83.0	85.3	-	-	-
MOST [25]	81.2	74.8	77.9	-	-	-
DeepSolo [49]	86.9	84.5	85.7	**64.2**	81.4	3.8
Proposed	**92.2**	84.4	**88.6**	59.3	79.1	1.7

Table 8: Comparative study on ICDAR2015 dataset for detection and spotting

Methods	Text Detection			Text Spotting (End-to-End)			Response Time
	P	R	F	S	W	G	
ABCNetv2 [45]	86.0	90.4	88.1	83.0	80.7	75.0	1.9
Huang et al. [4]	82.1	85.2	83.6	83.9	77.3	70.5	1.8
Zhang et al. [47]	90.3	89.7	90.0	85.2	79.4	73.6	**1.6**
Kittenplon et al. [48]	92.3	82.5	87.1	82.5	77.4	73.5	2.2
PAN++[1]	91.4	83.9	87.5	82.7	78.2	69.2	-
DBnet [22]	90.9	83.9	87.2	-	-	-	-
BRN-BCTS [14]	86.2	82.7	84.4	-	-	-	-
RCLM [15]	84.0	66.1	74.0	-	-	-	-
MaskTextSpotter v3 [41]	85.9	77.9	82.1	83.3	78.1	74.2	-
East [23]	34.4	34.7	34.5	-	-	-	-
SegLink [24]	73.1	76.8	75.0	-	-	-	-
MOST [25]	89.1	87.3	88.2	-	-	-	-
DeepSolo [49]	**92.4**	87.9	90.1	**88.1**	**83.9**	**79.5**	1.8
Proposed	89.7	**91.4**	**90.5**	83.9	81.1	75.5	1.7

The response times (seconds per batch of 16 images) of the proposed and the state-of-the-art methods are also given in Tables 5 to 8. One may have thought that by introducing the Hilbert transform, the proposed method would be slower. However, considering the response times for the two most complex datasets, the TTSD (Table 5) and the Total-Text dataset (Table 6), it is evident that the proposed method is more efficient than the state-of-the-art. The key reason for achieving such efficiency is due to the single network (GAN) designed for an end-to-end text spotting system. In addition, the proposed method does not depend much on the number of training samples available because the Hilbert transform step reduces the complexity of the problem, in contrast to the existing methods. However, for the CTW1500 and ICDAR2015 datasets, the proposed method is not the fastest, although it comes closely in second place, compared to the fastest state-of-the-art method. This indicates that for these two less complex datasets, the best state-of-the-art method is more efficient than the proposed one. Overall, however, the difference is very small and the proposed method is more efficient than most of the state-of-the art methods.

(a) Detection of challenging tattoo text by the proposed approach.
(b) Recognition results: "A flythe ….!@$(#&@," (left image), "XYLO&@#" (middle) and " @$(%@#($@@" (right image).

Fig. 15. Examples of correct detection but erroneous recognition results obtained by the proposed approach for complex tattoo text images.

D. *Limitations of the Proposed Method*

Naturally, there are some particularly challenging cases, such as those shown in Fig. 15, where the proposed approach fails to spot the tattoo text correctly. As discussed earlier, tattoo text detection and spotting are very challenging due to calligraphic writing, loss of character shapes, occlusion, overlap with the background design and skin deformation. When the shape and structure of characters is lost, the proposed method does not perform spotting well, as shown in the example failure cases in Fig. 15. The main reason is that individual character shapes and text structure are virtually absent. Furthermore, there is also scope for addressing other challenges of tattoo text spotting, such as the difficulty of identifying character shapes due to occlusion and calligraphic text. A promising solution may involve the combination of the proposed spotting model with a prediction model, such as a language model. Furthermore, there is always scope for extending the proposed method to further improve its performance on scene text spotting. One possibility is a language vision model integrated with a transformer.

V. Concluding Remarks

This paper has introduced a new end-to-end approach for spotting text in tattoo and natural scene images. The proposed approach benefits from a reduction in background complexity due to its use of the Hilbert transform resulting in more efficient candidate region detection. Accordingly, the performance of text detection and spotting increases. The generator and discriminator components of a GAN are subsequently used for text detection and recognition in an end-to-end fashion (spotting). Experimental results on a new tattoo text dataset (TTSD) and on existing benchmark scene text datasets show that the proposed approach outperforms the state-of-the-art methods in terms of detection and spotting of both tattoo and scene text. Future work will involve the exploration of language models as well as contextual knowledge from other text in the images to address the remaining challenges for spotting when tattoo text is not readable (i.e. the structure of the text is lost).

Acknowledgement

This project is partially funded by Technology Innovation Hub, Indian Statistical Institute, Kolkata, India.

REFERENCES

[1] W. Wang, E. Xie, X. Li, X. Liu, D. Liang, Z. Yang, T. Lu and C. Shen, "PAN++: Towards efficient and accurate end-to-end spotting of arbitrary-shaped text", IEEE Trans. PAMI, 2021. DOI:10.1109/TPAMI.2021.3077555.

[2] C. Zheng, Z. Wu, T. Wang, Y. Cai and Q. Li, "Object-aware multimodal named entity recognition in social media posts with adversarial learning", IEEE Transactions on Multimedia, pp 2520-2532, 2021.

[3] H. Zhang, S. Qian, Q. Fang and C. Xu, "Multimodal disentangled domain adaption for social media event rumor detection", IEEE Transactions on Multimedia, pp 4441-4454, 2021.

[4] M. Huang, Y. Liu, Y, Z. Peng, C. Liu, D. Lin, S. Zhu, N. Yuan, K. Ding, and L. Jin, "SwinTextSpotter: Scene Text Spotting via Better Synergy between Text Detection and Text Recognition", In Proc. CVPR, pp 4593-4603, 2022.

[5] P. N. Chowdhury. P. Shivakumara, R. Raghavendra, S. Nag, U. Pal, T. Lu and D. Lopresti, "An episodic learning network for text detection on human bodies in sports images", IEEE Trans. CSVT, Vol 32, pp. 2279-2289, 2021.

[6] S. Nag, P. Shivakumara, U. Pal, T. Lu and M. Blumenstein, "A new unified method for detecting text from marathon runners and sports players in video", PR, Vol. 107: 107476, 2020.

[7] H. Han, J. Li, A. K. Jain, S. Shan and X. Chen, "Tattoo image search at scale: Joint detection and compact representation learning", IEEE Trans. PAMI, 2333-2348, 2019.

[8] K. Wang, P. Xiao, X. Feng and G. Wu, "Image features detection from phase congruency based on two-dimensional Hilbert transform", PRL, 32, 2021.

[9] P. Keserwani and P. P. Roy, "Text region conditional generative adversarial network for text concealment in the wild", TCSVT, 2021.

[10] F. Zhan, H. Zhu and S. Lu, "Spatial fusion GAN for image synthesis", CVPR, 3648-3657, 2019.

[11] G. Deng, Y. Ming and J. H. Xue, "RFRN: A recurrent feature refinement network for accurate and efficient scene text detection", Neurocomputing, 465-481, 2021.

[12] Z. Raisi, M. A. Naiel, G. Younes, S. Wardell and J. S. Zelek, "Transformer-Based Text Detection in the Wild", CVPR, 3162-3171, 2021.

[13] P. Dai. Y. Li, H. Zhang, J. Li and X. Cao, "Accurate scene text detection via scale-aware data augmentation and shape similarity constraint", TMM, 2021.

[14] P. Dai, H. Zhang, and X. Cao, "Deep Multi-Scale Context Aware Feature Aggregation for Curved Scene Text Detection", IEEE Transactions on Multimedia, pp 1969-1984, 2020.

[15] P. Dai, S. Zhang, H. Zhang, and X. Cao, "Progressive contour regression for arbitrary-shape scene text detection", In Proc. CVPR, pp 7389-7398, 2021.

[16] Y. Wang, H. Xie, Z. Zha, M. Xing, Z. Fu and Y. Zhang, "ContourNet: Taking a further step toward accurate arbitrary-shaped scene text detection", CVPR, 1750-11759, 2020.

[17] M. Xue, P. Shivakumara, C. Zhang, Y. Xiao, T. Lu. U. Pal and D. Lopresti, "Arbitrarily-oriented text detection in low light natural scene images", TMM, 2021.

[18] S. X. Zhang, X. Zhu, J. B. Hou and C. Liu., "Deep relational reasoning graph network for arbitrary shape text detection", CVPR, 9696-9705, 2020.

[19] S. Zhang, Y. Liu, L. Jin, Z. Wei and C. Shen, "OPMP: An omnidirectional pyramid mask proposal network for arbitrary-shape scene text detection", TMM, 454-467, 2021.

[20] S. X. Zhang, Z. Xiaobin, Y. Chun W. Hongfa and X. C. Yin. "Adaptive Boundary Proposal Network for Arbitrary Shape Text Detection", ICCV, 1305-1314. 2021.

[21] Y. Zhu, C. Jianyong L. Lingyu K. Zhanghui L. Jin and W. Zhang. "Fourier contour embedding for arbitrary-shaped text detection", CVPR, 3123-3131. 2021.

[22] M. Liao, Z. Wan, C. Yao, K. Chen and X. Bai, "Real-time scene text detection with differentiable binarization", In Proc. AAAI, pp. 11474-11481, 2020.

[23] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He J. Liang, "East: an efficient and accurate scene text detector", In Proc. CVPR, pp 5551-5560, 2017. .

[24] B. Shi, X. Bai and S. Belongie, "Detecting oriented text in natural images by linking segments", In Proc. CVPR, pp 2550-2558, 2017.

[25] M. He, M. Liao, Z. Yang, H. Zhong, J. Tang, W. Cheng, C. Yao, Y. Wang and X. Bai, "MOST: A multi-oriented scene text detector with localization refinement", In Proc. CVPR, pp 8813-8822, 2021.

[26] L. Nandanwar, P. Shivakumara, R. Ramachandra, T. Lu, U. Pal, A. Antonacopoulos and Y. Lu, "A new deep wavefront based model for text localization in 3D video", IEEE Trans. CSVT, Vol. 32, pp. 3375-3389,2022.

[27] K. S. Raghunandan, P. Shivakumara, H. A. Jalab, R. W. Ibrahim, G. H. Kumar, U. Pal and T. Lu, "Riesz fractional based model for enhancing license plate detection and recognition", IEEE Trans. CSVT, pp 2276-2288, 2018.

[28] K. S. Raghunndan, P. Shivakumara, S. Roy, G. H. Kumar, U. Pal and T. Lu, "Multi-script-oriented text detection and recognition in videos/scene/born digital images", IEEE Trans. CSVT, pp 1145-1162, 2019.

[29] N. Lu, W. Yu, X. Qi, Y. Chen, P. Gong, R. Xiao and X. Bai, "Master: Multi-aspect non-local network for scene text recognition", PR, 117, 2021.

[30] Z. Qiao, X. Qin, Y. Zhou, F. Yang and W. Wang, "Gaussian Constrained Attention Network for Scene Text Recognition", ICPR, 3328-3335, 2021.

[31] P. Dai, H. Zhang and X. Cao, "SLOAN: Scale-Adaptive Orientation Attention Network for Scene Text Recognition", TIP, 1687-1701, 2021.

[32] Y. Gao, Y. Chen, J. Wang and H. Lu, "Semi-supervised scene text recognition", TIP, 3005-3016, 2021.

[33] Q. Lin, C, Luo, L. Jin and S. Lai, "STAN: A sequential transformation attention-based network for scene text recognition", PR, 111, 2021.

[34] R. Litman, O. Anschel, S. Tsiper, R. Litman, S. Mazor and R. Mamatha, "SCATTER: Selective context attentional scene text recognizer", CVPR, 11959-11969, 2020.

[35] C. Luo, Y. Zhu, L. Jin and Y. Wang, "Learn to augment: Joint data augmentation and network optimization for text recognition", CVPR, 13743-13752, 2020.

[36] Z. Qiao, Y. Zhou, D. Yang, Y. Zhou and W. Wang, "Seed: Semantics enhanced encoder-decoder framework for scene text recognition", CVPR, 13528-13537. 2020.

[37] U. Sajid, M. Chow, J. Zhang, T. Kim and G. Wang, "Parallel scale-wise attention network for effective scene text recognition", https://arxiv.org/abs/2104.12076v1, 2021.

[38] Z. Wan, J. Zhang, L. Zhang, J. Luo and C. Yao, "On vocabulary reliance in scene text recognition", CVPR, 11422-11431, 2020.

[39] C. Zhang, W. Ding, G. Peng, F. Fu and W. Wang, "Street view text recognition with deep learning for urban scene understanding in intelligent transportation systems", T-ITS, 4727-4743, 2021.

[40] C. Luo, L. Jin and J. Chen, "SimAN: Exploring self-supervised representation learning of scene text via similarity-aware normalization", In Proc. CVPR, 2022. DOI: 10.48550/arXiv.2203.10492

[41] M. Liao, G. Pang, J. Huang, T. Hassner and X. Bai, "Mask TextSpotter v3: Segmentation proposal network for robust scene text spotting", In Proc. ECCV, pp 706-722, 2020.

[42] H. Wang, P. Lu, H. Zhang, M. Yang, X. Bai, Y. Xu, M. He, Y. Wang and W. Lu, "All you need is boundary: Toward arbitrary-shaped text spotting", AAAI, 12160-12167, 2020.

[43] L. Qiao, S. Tang, Z. Cheng, Y. Xu, Y. Niu, S. Pu and F. Wu, "Text perceptron: Towards end-to-end arbitrary-shaped text spotting", In Proc. AAAI, pp 11899-11907, 2020.

[44] M. Liao, P. Lyu, M. He, C. Yao, W. Wu and X. Bai, "Mask TextSpotter: An end-to-end trainable neural network for spotting text with arbitrary shapes", IEEE Trans. PAMI, 532-548, 2021.

[45] Y. Liu, C. Shen, L. Jin, T. He, P. Chen, C. Liu and H. Chen, "ABCNetv2: Adaptive Bezier-curve network for real-time end-to-end text spotting", IEEE Trans. PAMI, 2021. DOI: 10.1109/TPAMI.2021.3107437

[46] P. Wang. H. Li and C. Shen, "Towards end-to-end text spotting in natural scenes", IEEE Trans. PAMI, 1-16, 2021. DOI: 10.1109/TPAMI.2021.3095916

[47] X. Zhang, Y. Su, S. Tripathi and Z. Tu, "Text Spotting Transformers", In Proc. CVPR, pp 9519-9528, 2022.

[48] Y. Kittenplon, I. Lavi, S. Fogel, Y. Bar, R. Manmatha and P. Perona, "Towards Weakly-Supervised Text Spotting using a Multi-Task Transformer", In Proc. CVPR, pp 4604-4613, 2022.

[49] M. Ye, J. Zhang, S. Zhao, J. Liu, T. Liu, B. Du and D. Tao, "Deepsolo: Let transformer decoder with explicit points solo for text spotting", In Proc. CVPR, pp 19348-19357, 2023.

[50] Zhan, H. Zhu and S. Lu, "Spatial fusion GAN for image synthesis", In Proc. CVPR, pp 3468-3657, 2019.

[51] D. M. Souza, J. Wehrmann and D. D. Ruiz, “Efficient neural architecture for text-to-image synthesis “, In Proc. IJCNN, 2020. DOI:10.1109/IJCNN48605.2020.9207584
[52] C. Liu, Y. Liu, L. Jin, S. Zhang, C. Luo and Y. Wang, “EraseNet: End-to-End text removal in the wild”, IEEE Trans. Image Processing, pp 8760-8775, 2020.
[53] Y. Li, Q. Yan, Y. Huang and L. Gao, “A GAN-based feature generator for table detection”, In Proc. ICDAR, pp 763-768, 2019.
[54] T. Hinz, S. Heinrich and S. Wermter, “Semantic obeject accuracy for generative text-to-image synthesis”, IEEE Trans. Pattern Analysis and Mchine Intelligence, pp 1552-1565, 2022.
[55] P. Dzianok, M. Koldziej and E. Kublik, “Detecting attention in Hilbert transformed EEG brain signals from simple reaction and choice reaction cognitive tasks”, In Proc. BIBE, 2021. DOI: 10.1109/BIBE52308.2021.9635187
[56] A. Dragulinescu, A. M. Dragulinescu and I. Marcu, “Optical correlator based on Hilbert transform for image recognition”, In Proc. EACI, pp.1-4; 2021. DOI: 10.1109/ECAI52376.2021.9515064
[57] M. Zabin, J. Uddin, H. J. Choi, M. H. Furthad and A. B. Ullah, “Industrial fault diagnosis using Hilbrert transform and texture features”, In Proc. BigComp, pp 121-128, 2021.
[58] T. Chowdhury, P. Shivakumara, U. Pal, T. Lu and R. Ramachandra and S. Chanda, "DCINN: Deformable Convolution and Inception Based Neural Network for Tattoo Text Detection through Skin Region”, ICDAR, 335-350, 2021.
[59] R. Jiang, “Riesz transform via heat kernel and harmonic functions on non-compact manifolds”, Advances in Mathematics, 377, 2021.
[60] J. Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang and E. Shechtman, “Toward multimodal image-to-image translation”, Advances in neural information processing systems, 30, 2017.
[61] B. Liu, K. Song, Y. Zhu, G. de Melo and A. Elgammal, “Time: text and image mutual-translation adversarial networks”, In Proc. AAAI, pp 2082-2090, 2021.
[62] B. Duan, W. Wang, H. Tang, H. Latapie and Y. Yan, “Cascade attention guided residue learning gan for cross-modal translation”, In Proc. ICPR pp 1336-1343, 2021.
[63] T. Xu and W. Takano, “Graph Stacked Hourglass Networks for 3D Human Pose Estimation”, CVPR, 16105-16114, 2021.
[64] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai and L. Van Gool, "Multi-Task Learning for Dense Prediction Tasks: A Survey," IEEE Trans. PAMI, 2021.
[65] S. Kwon, “Att-Net: Enhanced emotion recognition system using lightweight self-attention module”, Applied Soft Computing, 2021.
[66] K. Li, Z. Fu, H. Wang, Z. Chen and Y. Guo, “Adv-Depth: Self-Supervised Monocular Depth Estimation With an Adversarial Loss”, IEEE Signal Processing Letters, 38-642, 2021.
[67] H. H. Lee, S. K. Ko and Y. S. Han, “SALNet: Semi-supervised few-shot text classification with attention-based lexicon construction”, In Proc. AAAI, pp. 13189–13197, 2021.
[68] Y. Liu, L. Jin, S. Zhang and S. Zhang, “Detecting curve text in the wild: New dataset and new solution”, https://arxiv.org/abs/1712.02170, https://doi.org/10.48550/arXiv.1712.02170.
[69] C. K. Ch’ng and C. S. Chan, “Total-Text: A comprehensive dataset for scene text detection and recognition”, In Proc. ICDAR, pp 935-942, 2017.
[70] D. Karatzas *et al.*, "ICDAR 2015 competition on Robust Reading”, In Proc. ICDAR, pp 1156-1160, 2015. doi: 10.1109/ICDAR.2015.7333942.
[71] D. Li, H. Ling, S. W. Kim, K. Kreis, S. Fidler and A. Torralba, “BigDatasetGAN: Synthesizing ImageNet with Pixel-wise Annotations”, In Proc. CVPR, pp 21330-21340, 2022.