Domain Adaptive Scene Text Localization in Night View Images

Majed Alowaidi, Mohammed Alshehri
Department of Information Technology
College of Computer and Information Sciences
Majmaah University, Majmaah-11952, Saudi Arabia
Email ID: m.alowaidi@mu.edu.sa, ma.alshehri@mu.edu.sa

***Abstract*—Scene text localisation in a night is a challenging problem and a less explored problem in the area of robust reading. The day and night images have a different distribution. Hence, a domain shift exists from day to night images. This paper introduces a domain adaptation network (DAN) to learn the domain shift from day images to night images. The proposed sub-network DAN can be attached to any existing state-of-the-art text detector to learn the domain shift from day to night. For the training of the DAN, synthetic data generation has been done by utilising the generative adversarial network. The proposed method has been extensively validated on the public as well as on the synthetic datasets. The proposed DAN improves the performance of the underline text localisation model with a margin of 1.4-4.0%, 0.1-1.0%, and 0.1-3.1% on LP Night Dataset, ICDAR2015, and Total-Text benchmarking datasets.**

***Index Terms*—Scene Text Localisation, Domain Adaptation, Night images, Deep learning**

I. Introduction

Text localisation [1], [2] is one of the hottest topics among the computer vision and pattern recognition research community. The text localisation task is to generate a bounding box for the text instances in the natural images [3]. Text in the natural images may appear on the signboard, shop front, license plate, t-shirts, etc. The localisation of the text in the natural image is difficult due to the complex background and omnipresence of text instances. Various applications are dependent on efficient text localisation on natural images. Some essential applications of text localisation are automatic image tagging, robot navigation, and text reading for a visually impaired person [4].

Scene text presence on certain surfaces is more probable than other surfaces [5]. For example, the presence of text is more probable on signboards, car license plates, etc. but very rare on a leaf of a tree, on cloud and sea surface. This probable presence of text on certain surfaces helps the existing method learn and emphasises more importance on those surfaces. However, the various range of text scale, aspect ratio, and lighting conditions make this problem challenging [1]. Most of the existing work [6]–[9] mainly targets the challenge such as complex background and textual properties such as scale, orientation, and aspect ratio. But, only a few of the existing work [10]–[13] focuses on the impact of the lighting condition on text detection.

Fig. 1: The sample images of LP Night dataset [19] to show the illumination issue and the submerged text with background.

In the past, the research on scene text localisation was considered images with good lighting conditions. Even the most popular scene text localization public datasets such as ICDAR 2015 [14], MLT2019 [15], and COCO-Text [16] have considered the day images. The night view images remain very less explored by the research community. Text localisation in the night is also a significant problem to get solved. The efficient automatic text detection at night is very beneficial to help persons suffering from night blindness [17] and unable to see at night. It has become also helpful in intelligent Driving Assistant System [18] at night time. In this work, we focused on the text localisation method, which efficiently handles the lightning condition for night images.

There are two important challenges for text localisation in night images. The first challenge is the lack of well-labelled night images to train the text localisation method. The deep learning-based text localisation method needs a vast amount of well-labelled training data. But, all the existing big dataset, such as COCO-Text [16] and MLT2019 [15], consists of day images. Hence, the need for night images for the training of the deep neural network for the scene text localisation at night

(a) Day image

(b) Night image

Fig. 2: Performance of state-of-the-art method DB [6] trained on day images and tested on (a) Day image, and (b) Night image.

is one of the fundamental challenges.

The second challenge is the difficulty of the appearance of text in an image due to bad illumination during nighttime. Some text instances are deeply submerged with the background due to poor illumination. Few sample images to show the above-mentioned challenges in night images are shown in Fig. 1. Due to the domain shift, the model trained on day images is not performing well in the night images. Figure 2 shows the performance of the state-of-the-art method trained on day images and text on the day and night images. The data distribution of the day images is different from the night images. Hence, a domain adaptation network is needed to learn the domain shift from day to night. Including domain shift network over the existing state-of-the-art methods can help to improve their performance on the night images.

The major contributions of the work are as follows:

1) Proposed a domain-adaption network that helps to improve the performance of the existing state-of-the-art method for text localisation in the night images.
2) Synthetic data generation has been done to generate night view images that help train the deep learning-based scene text localiser.
3) Extensive experimental validation of the proposed method has been done on the public and the synthetic data.

The rest of the paper is organised into four sections. The work related to this work is discussed in Section II. The proposed method has been discussed in detail in Section III. The experiments and results are described in Section IV and Section V respectively.

II. Related Work

Text localisation in the natural images has gained a lot of attention in the last decade by the research community with using machine learning [1], [2]. After the advent of deep learning in the area of text localisation, there is a number of works in which achieved a good performance in natural images for various publicly available text localisation datasets [6], [8], [20]. The text localisation on natural images is broadly classified into two categories based on lightning conditions: day and night images.

A. Text Localization in Day Images

The text localisation for the day images is the main centre of attraction among researchers of computer vision. The various competitions that have been launched under the name of the robust reading competition show the popularity of the text localisation [14], [15], [21]. All these competitions are mainly focused on the day images. The methods that solved these day majority images datasets are based on regression-based approach, segmentation-based approach, or both. The regression can be done in three ways, namely, direct regression [22], indirect regression [7], and combination of direct and indirect regression [8]. The regression-based approaches [7], [20] are based on the state-of-the-art object detectors such as YOLO [23] and Faster-RCNN [24] . These text detectors are based on the region proposals. Based on the anchor boxes, various bounding boxes are generated from the image grid. Various aspect ratio anchor boxes have been taken for each image grid to improve the recall. The other approach is the segmentation-based scene text detection [6], [25], [26] in night images. These methods have segmented the whole image into categories such as text, background, and boundary pixels. These methods are based on the feature pyramid network [27]. The feature pyramid network helps to capture the text of various scales in the images. The third approach is the combination of the regression and segmentation [8], [28], [29]. The bounding box regression has been done for each image grid, and the corresponding box scores have been obtained from the semantic segmentation. All these approaches have not been designed to handle the domain shift issue of night images. In this work, we have proposed a subnetwork, i.e., domain adaptation network, that handles this domain shift and can be included as a sub-network among all the above state-of-the-art approaches to get good performance in day images and get adopted for the night images.

B. Text Localization in Night Images

For the text localisation at night, there are only a few words that appear in the past [12], [13], [30]. In [30], has proposed a localisation method for both day and night images both. The method is based on adaptive thresholding on a grayscale image. In [12], the authors have proposed a method that is based on gradient vector flow and also introduced a method of augmentation to fuse the gradient of red, green, and blue channels for determining the dominant pixels. The gradient vector flow has been applied over the fused information for text extraction in the images. The above-mentioned methods are not based on deep learning. These handcraft feature-based methods are not robust enough to be used in various night scenarios. In [13], the authors have proposed a deep learning-based method for localising the license plate at low lightning images and also extending its work for the night images. They have enhanced the pixels of the image to handle the localisation of the text in license plate at low-light as well as at night images. They combine the popular semantic segmentation method known as UNet [31] and EAST text detector [28]. The UNet has been used for the enhancement of

the image, whereas the text detector is acting as an adversarial network for the enhancement of the input image.

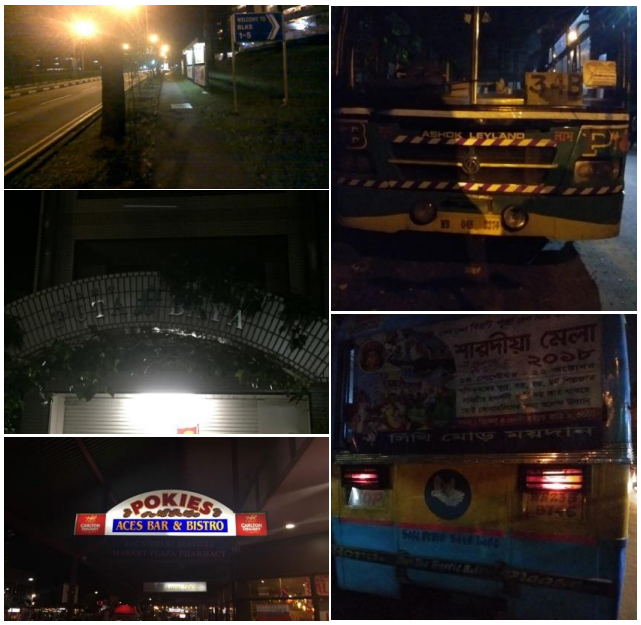

Fig. 3: Some Samples exemplifying the issues of night view images from ICDAR15, TotalText, and LP datasets.

III. Proposed Methodology

A. Overview of the work

The night view images have very low-intensity pixels as the majority, and if some light source is present, then light distribution is also uneven in its surrounding. The proposed domain adaptation network (DAN) as a sub-network combined with the existing scene text localisation models is responsible for adjusting the input image, its mean, and scale (range). This adjustment is spatially local for each pixel, thus behaving differently for low-intensity and light-reflecting regions 3. The whole text localisation model (DAN+text localisation model) learn the local mean and scale at each pixel by DAN, a separate four-layer convolution neural network (refer section III-B). The overview of the proposed text localisation is depicted in Fig. 5. Here, the proposed domain adaptation network first processes the input image, which yields the mean and scale vector map as the output. These mean and scale vector maps with the original input image are used to obtain the adjusted image by equation 1. This modified image is then processed by a scene text localisation method to generate text proposals.

B. Domain Adaptation Network (DAN)

The domain adaptation network (DAN) is a four-layered convolution neural network. The architecture of the network is depicted in Fig. 6, and the configuration detail of each layer of DAN is tabulated in Table I. All the convolution layers used in DAN are processed with stride 1 and dilation 1. The initial two layers C_1^1, C_1^2 has no bias term as batch normalisation layers follow them, whereas the C_1^3, C_1^4 layers has the bias term. All the trainable parameters of DAN initialised with HeNormal [32] initialiser. The proposed Domain Adaptation Network generates the mean vector and scale vector at each pixel location based on the pixel's neighbourhood. These mean and scale vectors are used to normalise the input image by equation 1. Where $I^n(x,y)$ and $I^0(x,y$ are the generated normalized image and original image at location x,y, and $V_{mean}(x,y)$ and $V_{scale}(x,y)$ are the mean and scale vector at x,y location.

$$I^n(x,y) = V_{scale}(x,y) \times (I^o(x,y) - V_{mean}(x,y)) \quad (1)$$

It seems that the working of DAN is similar to a batch-normalisation layer applied on the input images. But batch-normalisation layer has a single mean and scale vector, whereas the proposed DAN produced a different mean and scale vector at each pixel location of the input image. Thus, DAN's normalisation is highly dependent on the spatial locality of the underline pixel location. This spatial dependency of the generated normalised image better represents the loss function minimisation in the training phase.

TABLE I: The architecture description of the Domain Adaptation Network.

Layer	#Kernel	KernelSize	#Parameter	OutputSize
Input	0	0	0	$H \times W \times 3$
C_1^1	32	7×7	4704	$H \times W \times 32$
BatchNorm	32	-	64	$H \times W \times 32$
C_1^2	32	3×3	9216	$H \times W \times 32$
BatchNorm	32	-	64	$H \times W \times 32$
C_1^3	3	1×1	99	$H \times W \times 3$
C_1^4	3	1×1	99	$H \times W \times 3$
Total number of parameter in Domain Adaptation Network: **14246**				

C. Text Localization Model

The proposed domain adaptation network (DAN) can normalise the input data according to the optimisation used in the training phase. An experimental study is done to show the effectiveness of the DAN with some existing scene text detection methods. In this study we are utilized some current state-of-the-art methods as Differential Binarization (DB) [6], TextFuseNet [9], and TextMountain [25]. The training procedure used to train the whole model (DAN + text localisation model) is the same as the original text localisation model without DAN as a sub-network. Please refer to the corresponding method for the detail of the text localisation model and its hyperparameter settings.

IV. Experimental Setup

A. Dataset Used

a) Synthetic Data: A large amount of labelled data is required to train a deep neural network by a supervised machine learning approach. But the number of labelled scene text images with night view in existing datasets is very low.

Fig. 4: The Qualitative Results Obtained by Differential Binarization (DB) [6] method without and with Proposed Domain Adaptation Network (DAN). Here top row corresponds to the original image, and the bottom is for night view images. The left column is the input images, mid column obtained by DB method without DAN, and the right column results obtained by DB with DAN.

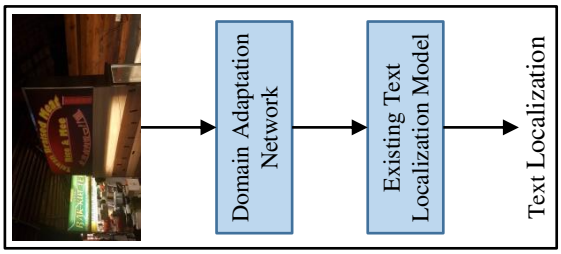

Fig. 5: Overview of the Proposed Method.

Therefore we are generating these training images synthetically. We used the CycleGAN approach to convert the scene text images of conventional datasets [ICDAR 2015 [14], Total Text [33]] into their corresponding night images.

b) ICDAR2015: ICDAR2015 (IC15) [14] dataset is provided for challenge4 ICDAR2015 Robust Reading Competition. The dataset provides the word level annotation. Besides this, texts that are unclear (not readable) or small are masked as "DONOTCARE". This dataset provides 1000 images for training and 500 for testing.

c) Total-Text: Total-text dataset [33] provide a range of text orientation and shapes for scene text detection. This dataset consists of multi-oriented and curved shape text with word-level annotation. This dataset provides 1255 images for training and 300 for testing.

d) LP Night Dataset: LP Night Dataset [19] is proposed for the evaluation purpose of license plate detection in low light images. The dataset is created by capturing license plate images at nights having low light and limited light conditions. This dataset is constituted of 200 images with low contrast, poor quality, affected lights, multiple headlights from different vehicles, etc.

B. Evaluation Criteria

The performance analysis of the proposed method is done with three different measurements. These measurements are Recall, Precision, F-measure. The evaluation of text localisation relies on the $precision$ (P) and $recall$. Generally, a text localisation method uses some threshold to decide a text proposal is valid or not. The precision and recall vary with this threshold. Here, the recall of the method can be improved by decreasing the threshold while precision diminishes. Therefore, another measure $f-measure$ is adopted to counter the trade-off between $precision$ and $recall$ and soften the threshold selection effect. The $f-measure$ is obtained by equation 2, where TP is true-positive, FP is false-positive, and FN is false-negative. A predicted text proposal is considered correct if its score is greater than a predefined threshold. We used the same threshold as the base method (Differential Binarization (DB) [6], TextFuseNet [9], and TextMountain [25]).

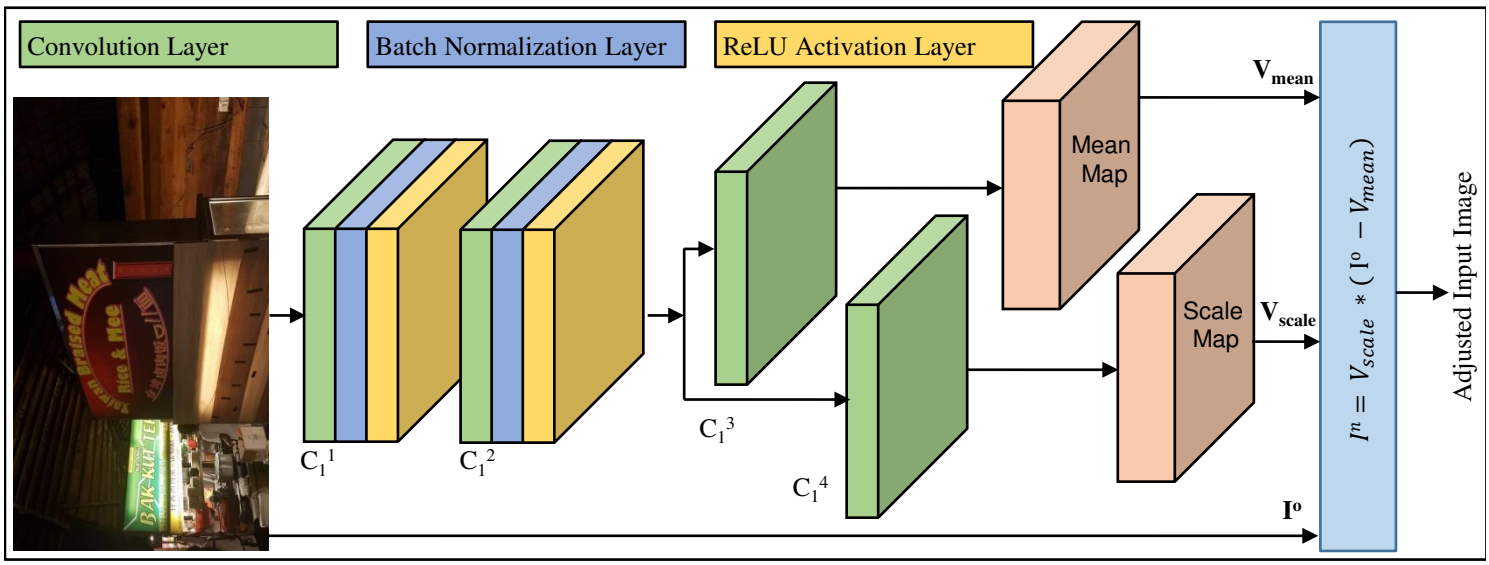

Fig. 6: The Architecture Description of the proposed Domain Adaptation Network.

$$
\begin{aligned}
precision \quad & P = \frac{TP}{TP + FP} \\
recall \quad & R = \frac{TP}{TP + FN} \\
f - measure \quad & F = \frac{2 \times P \times R}{P + R}
\end{aligned} \tag{2}
$$

V. Results and Analysis

Fig. 4 shows the qualitative results of the proposed domain adaptation network to enhance both day and night images. It indicates that the DAN improved the text localisation proposals of the Differential Binarization (DB) [6] method in both day and night images. Besides this, it also shows that the DAN is effectively normalised the input image to get better performance for the text localisation task.

Table II,III,and IV tabulates the performance results of different methods with and without the proposed domain adaptation network (DAN). The methods are trained and tested with different versions (original images and synthetically generated night images by CycleGAN). Here all the images are resized such that the height of the images becomes 800 pixels.

The Quantitative results over the LP Night Dataset [19] are tabulated in Table II. Real images constitute this dataset with the low-lighting and variable lighting conditions at night. We have trained different methods without and with DAN as a sub-network and tabulated the result. The performance of all the methods is improved with a margin of 1.4-4.0% when they are trained with the DAN.

Table III illustrates the performance measures of different methods with different variations in training and testing data of ICDAR 2015 (IC15) [14] dataset. This table shows that if training and testing data are from different domains (day and night images), then the performance of the underline methods decreases drastically. The loss in the performance is regained if the method is trained with corresponding images, but it still lacks the performance achieved in less complex data. This difference indicates that the methods are not able to adjust to variations in the input data. Therefore when the DAN is added as a sub-network to these methods, their performance improves (margin of 0.1-1.0%) again.

TABLE II: Text Detection performance comparison on the LP Night Dataset [19]. The blue and red color shows the best and second-best performance in the table. Here DAN stands for the domain adaptation network.

Method	Performance Training: Original Data Testing: Original Data Method: without DAN			Performance Training: Original Data Testing: Original Data Method: with DAN		
	Precision	Recall	F-measure	Precision	Recall	F-measure
TextMountain [25]						
VGG16	88.2	75.6	81.4	88.1	78.8	83.2
ResNet50	87.8	74.8	80.8	87.9	77.2	82.2
DB [6]						
ResNet18	87.5	70.5	78.1	87.3	77.1	81.9
ResNet50	88.1	72.3	79.4	88.4	79.0	83.4
TextFuseNet [9]						
ResNet50	88.3	76.4	81.9	88.1	80.6	84.2
ResNet101	88.9	78.2	83.2	88.7	83.2	85.9

Table IV illustrates the performance measures over the Total-text dataset [33] which provide a range of text orientation and shapes for scene text detection. The variation in training and testing images are the same as the performance analysis of the ICDAR 2015 dataset. Similar to the ICDAR 2015, the methods with DAN yield the best performance with a margin of 0.1-3.1% in all settings.

TABLE III: Text Detection performance comparison on the ICDAR2015 [14] dataset. The blue and red color shows the best and second-best performance in the table. Here DAN stands for the domain adaptation network.

Method	Performance Training: Original Data Testing: Original Data Method: without DAN			Performance Training: Original Data Testing: Night Data Method: without DAN			Performance Training: Night Data Testing: Night Data Method: without DAN			Performance Training: Night Data Testing: Night Data Method: with DAN		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
TextMountain [25]												
VGG16	88.9	81.7	85.1	88.8	76.7	82.3	89.1	80.3	84.5	89.2	80.8	84.8
ResNet50	86.4	83.1	84.7	86.6	75.1	80.4	86.8	79.7	83.1	86.8	81.3	84.0
DB [6]												
ResNet18	84.7	76.9	80.6	84.8	74.8	79.5	85.1	76.0	80.3	85.0	76.3	80.4
ResNet50	87.2	82.9	85.0	87.4	75.6	81.1	87.6	80.2	83.7	87.5	81.2	84.2
TextFuseNet [9]												
ResNet50	90.1	86.8	88.4	89.2	76.9	82.6	89.3	82.4	85.7	89.7	84.0	86.8
ResNet101	94.6	88.6	91.5	93.8	79.7	86.2	94.2	84.9	89.3	94.4	86.5	90.3

TABLE IV: Text Detection performance comparison on the Total-Text [33] dataset. The blue and red color shows the best and second-best performance in the table. Here DAN stands for the domain adaptation network.

Method	Performance Training: Original Data Testing: Original Data Method: without DAN			Performance Training: Original Data Testing: Night Data Method: without DAN			Performance Training: Night Data Testing: Night Data Method: without DAN			Performance Training: Night Data Testing: Night Data Method: with DAN		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
TextMountain [25]												
VGG16	88.6	78.8	83.4	89.2	73.4	80.5	88.7	79.4	83.8	88.6	79.8	84.0
ResNet50	88.1	82.4	85.2	88.7	72.7	80.0	88.3	78.6	83.2	88.6	79.4	83.7
DB [6]												
ResNet18	88.3	77.9	82.8	89.0	72.5	79.9	88.6	77.8	82.8	90.1	77.9	83.6
ResNet50	97.1	82.5	89.2	88.3	73.1	80.0	92.6	78.1	84.7	94.3	82.3	87.9
TextFuseNet [9]												
ResNet50	87.5	83.2	85.3	89.4	75.8	82.0	88.4	80.4	84.2	91.2	82.7	86.7
ResNet101	89.0	85.3	87.1	95.2	77.5	85.4	92.0	81.7	86.5	93.2	86.5	89.7

VI. CONCLUSION

In this work, we have proposed a domain adaptive network as a sub-network that adjust (normalise) the input images and improve the performance of the underline text localisation method. We have also created different night view images corresponding to the real dataset ICDAR 2015 (IC15) [14] and Total-Text [33] with CycleGAN. Besides this, the performance analysis of the proposed domain adaptive network is also executed on real (LP Night Dataset [19]) and synthetically generated images (for IC15 and Total-Text dataset). The performance analysis study shows that the text localisation model's performance improves significantly with the proposed DAN. The text localisation results obtained with DAN in night view images are even better than the corresponding original day images, which indicates the performance enhancement obtained by the DAN over the original text localisation model.

REFERENCES

[1] Qixiang Ye and David Doermann, "Text detection and recognition in imagery: A survey," *IEEE Transactions on Pattern Analysis and Machine intelligence*, vol. 37, no. 7, pp. 1480–1500, 2014.

[2] Xiyan Liu, Gaofeng Meng, and Chunhong Pan, "Scene text detection and recognition with advances in deep learning: a survey," *International Journal on Document Analysis and Recognition*, vol. 22, no. 2, pp. 143–162, 2019.

[3] Xu-Cheng Yin, Ze-Yu Zuo, Shu Tian, and Cheng-Lin Liu, "Text detection, tracking and recognition in video: a comprehensive survey," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2752–2773, 2016.

[4] Yingying Zhu, Cong Yao, and Xiang Bai, "Scene text detection and recognition: Recent advances and future trends," *Frontiers of Computer Science*, vol. 10, no. 1, pp. 19–36, 2016.

[5] Anna Zhu, Renwu Gao, and Seiichi Uchida, "Could scene context be beneficial for scene text detection?," *Pattern Recognition*, vol. 58, pp. 204–215, 2016.

[6] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai, "Real-time scene text detection with differentiable binarization.," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 11474–11481.

[7] Minghui Liao, Baoguang Shi, and Xiang Bai, "TextBoxes++: A single-shot oriented scene text detector," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3676–3690, 2018.

[8] Prateek Keserwani, Ankit Dhankhar, Rajkumar Saini, and Partha Pratim Roy, "Quadbox: Quadrilateral bounding box based scene text detection using vector regression," *IEEE Access*, vol. 9, pp. 36802–36818, 2021.

[9] Jian Ye, Zhe Chen, Juhua Liu, and Bo Du, "Textfusenet: Scene text detection with richer fused features.," in *IJCAI*, 2020, pp. 516–522.

[10] Sabyasachi Mohanty, Tanima Dutta, and Hari Prabhat Gupta, "An efficient system for hazy scene text detection using a deep cnn and patch-nms," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2588–2593.

[11] Chaitanya Animesh, Sabyasachi Mohanty, Tanima Dutta, and Hari Prabhat Gupta, "Fast text detection from single hazy image using smart device," in *2017 IEEE International Conference on Multimedia & Expo Workshops*. IEEE, 2017, pp. 423–428.

[12] Pinaki Nath Chowdhury, Palaiahnakote Shivakumara, Umapada Pal, Tong Lu, and Michael Blumenstein, "A new augmentation-based method for text detection in night and day license plate images," *Multimedia Tools and Applications*, vol. 79, no. 43, pp. 33303–33330, 2020.

[13] Pinaki Nath Chowdhury, Palaiahnakote Shivakumara, Ramachandra Raghavendra, Umapada Pal, Tong Lu, and Michael Blumenstein, "A new u-net based license plate enhancement model in night and day images," in *Asian Conference on Pattern Recognition*. Springer, Cham, 2019, pp. 749–763.

[14] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al., "Icdar 2015 competition on robust reading," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2015, pp. 1156–1160.

[15] Nibal Nayef, Yash Patel, Michal Busta, Pinaki Nath Chowdhury, Dimosthenis Karatzas, Wafa Khlif, Jiri Matas, Umapada Pal, Jean-Christophe Burie, Cheng-lin Liu, et al., "Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 1582–1587.

[16] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie, "Coco-text: Dataset and benchmark for text detection and recognition in natural images," *arXiv preprint arXiv:1601.07140*, 2016.

[17] Alfred Sommer, Gusti Hussaini, Ignatius Tarwotjo, Djoko Susanto, and J Sulianti Saroso, "History of nightblindness: a simple tool for xerophthalmia screening," *The American journal of clinical nutrition*, vol. 33, no. 4, pp. 887–891, 1980.

[18] Sangeeth Reddy, Minesh Mathew, Lluis Gomez, Marçal Rusinol, Dimosthenis Karatzas, and CV Jawahar, "Roadtext-1k: Text detection & recognition dataset for driving videos," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 11074–11080.

[19] Pinaki Nath Chowdhury, Palaiahnakote Shivakumara, Ramachandra Raghavendra, Umapada Pal, Tong Lu, and Michael Blumenstein, "A new u-net based license plate enhancement model in night and day images," 2019.

[20] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu, "TextBoxes: A fast text detector with a single deep neural network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, vol. 31.

[21] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras, "ICDAR 2013 robust reading competition," in *12th International Conference on Document Analysis and Recognition*. IEEE, 2013, pp. 1484–1493.

[22] Wenhao He, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu, "Deep direct regression for multi-oriented scene text detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 745–753.

[23] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[25] Yixing Zhu and Jun Du, "Textmountain: Accurate scene text detection via instance segmentation," *Pattern Recognition*, vol. 110, pp. 107336, 2021.

[26] Oshada Jayasinghe, Sahan Hemachandra, Damith Anhettigama, Shenali Kariyawasam, Ranga Rodrigo, and Peshala Jayasekara, "Ceymo: See more on roads-a novel benchmark dataset for road marking detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3104–3113.

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[28] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang, "East: an efficient and accurate scene text detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5551–5560.

[29] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao, "TextSnake: A flexible representation for detecting text of arbitrary shapes," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 20–36.

[30] Rahim Panahi and Iman Gholampour, "Accurate detection and recognition of dirty vehicle plate numbers for high-speed applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 4, pp. 767–779, 2016.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[33] Chee Kheng Ch'ng and Chee Seng Chan, "Total-Text: A comprehensive dataset for scene text detection and recognition," in *14th IAPR International Conference on Document Analysis and Recognition)*. IEEE, 2017, vol. 1, pp. 935–942.