

Article

Towards Accurate Scene Text Detection with Bidirectional Feature Pyramid Network

Dongping Cao ^{1,2}, Jiachen Dang ^{1,2} and Yong Zhong ^{1,2,*}

¹ Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu 610041, China; caodongping16@mailsucas.ac.cn (D.C.); dangjiachen18@mailsucas.ac.cn (J.D.)

² School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China

* Correspondence: zhongyong@casit.com.cn

Abstract: Scene text detection, this task of detecting text from real images, is a hot research topic in the machine vision community. Most of the current research is based on an anchor box. These methods are complex in model design and time-consuming to train. In this paper, we propose a new Fully Convolutional One-Stage Object Detection (FCOS)-based text detection method that can robustly detect multioriented and multilingual text from natural scene images in a per pixel prediction approach. Our proposed text detector employs an anchor-free approach, unlike state-of-the-art text detectors that do not rely on a predefined anchor box. In order to enhance the feature representation ability of FCOS for text detection tasks, we apply the Bidirectional Feature Pyramid Network (BiFPN) as the backbone network, enhancing the model learning capacity and increasing the receptive field. We demonstrate the superior performance of our method on multioriented (ICDAR-2015, ICDAR-2017 MLT) and horizontal (ICDAR-2013) text detection benchmark tasks. Moreover, our method has an f-measure of 88.65 and 86.32 for the benchmark datasets ICDAR 2013 and ICDAR 2015, respectively, and 80.75 for the ICDAR-2017 MLT dataset.

Keywords: scene text detection; multioriented text; convolutional neural networks



Citation: Cao, D.; Dang, J.; Zhong, Y. Towards Accurate Scene Text Detection with Bidirectional Feature Pyramid Network. *Symmetry* **2021**, *13*, 486. <https://doi.org/10.3390/sym13030486>

Academic Editor: Katsumi Sakata

Received: 24 February 2021

Accepted: 7 March 2021

Published: 16 March 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Scene text detection is both fundamental and challenging in the field of machine vision, and plays a critical role in subsequent text recognition tasks. Current mainstream text detection methods [1–5], rely on predefined anchor frames to extract high-quality word candidate regions. Despite the success of these methods, they are associated with the following limitations: (1) text detection results are highly sensitive to the size, orientation, and the number of predefined anchor boxes. (2) Due to the variable size, shape, and orientation of text in natural scenes, it is difficult to capture all the text instances via the predefined anchor boxes. (3) A large number of anchor boxes are required in order to improve text detection performance, resulting in complex and time-consuming calculations. For example, in order to improve the accuracy of DeepText [1], Zhong empirically designed four scales and six aspect ratios, resulting in 24 prior bounding boxes at each sliding position. The number of anchors is 2.6 times more than those of Faster R-CNN [6].

Fully convolutional networks (FCNs) [7] recently have been very successful in the dense prediction task, including semantic segmentation [8–10], keypoint detection [11,12], and depth estimation [13,14]. Several scene text detection methods [15–19] treat text detection as a semantic segmentation problem and use the FCN for pixel-level text prediction. Liao et al. [20] proposed a novel binarization module called differentiable binarization (DB), which enabled the segmentation network to set the threshold of binarization adaptively, greatly improving the performance of text detection. Such methods typically diverge from anchor boxes to anchor-free frameworks by using corner/center points. This facilitates computational efficiency and generally improves the performance over anchor box-based

text detectors. Since only coarse text blocks can be detected from the saliency map, a complex postprocessing step is required to extract the precise bounding boxes [21]. Recently, Tian et al. [22] proposed a fully convolutional one-stage object detector (FCOS) pipeline to solve this issue and achieve state-of-the-art results in general object detection tasks.

In this paper, we design a simple yet efficient anchor-free method based on FCOS, a one-stage fully convolutional text detection framework with a weighted bidirectional pyramid feature network (BiFPN) for the scene text detection of natural images. Our proposed method's architecture is shown in Figure 1. In order to enhance the feature representation ability, we employ EfficientNet as a new backbone network. Experiments demonstrate the superior performance of our proposed approach compared to state-of-the-art methods for benchmark (ICDAR-2013, ICDAR-2015, ICDAR-2017 MLT) text detection datasets.

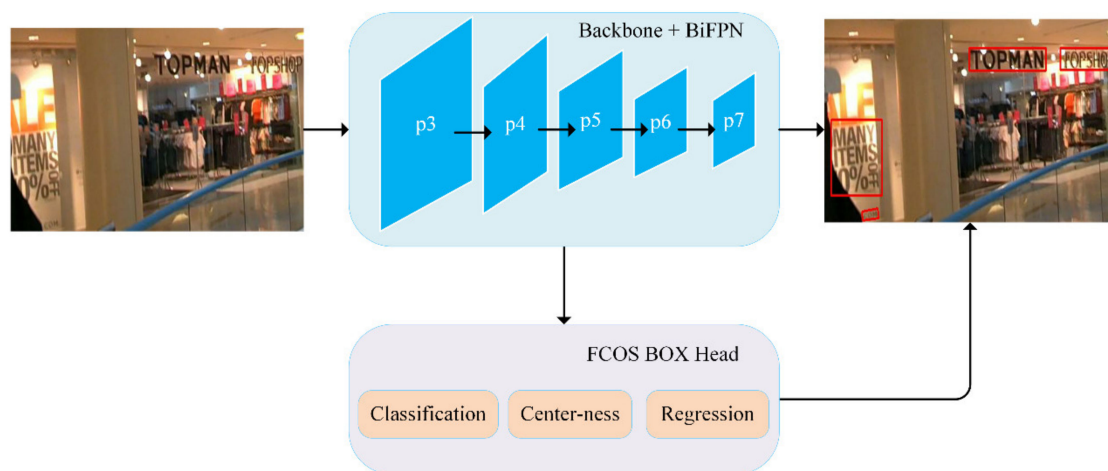


Figure 1. Architecture of the proposed fully convolutional one-stage (FCOS)-based text detector, comprising a backbone, bidirectional feature pyramid network (BiFPN) and FCOS box head for text detection.

2. Related Work

Anchor-based text detector. Text detection methods based on region proposals use a general object detection framework and often employ regression text boxes to get the region text information [23]. For example, in [1], the GoogleNet [24] inception structure was employed to improve Faster R-CNN [25]. As a result, an initial region proposal network (InceptionRPN) was generated, which acquired text candidate regions, removed background regions using a text detection network, and voted on the detected overlapping regions to determine the optimal result. Jiang et al. [26] proposed a rotational region convolutional network (R2CNN) to detect arbitrarily-oriented text in scene images. A novel connectionist text proposal network (CTPN) was proposed in [27] in order to locate text lines in scene images. In [28], a vertically regressed proposal network (VRPN) was proposed to match text regions using multiple neighboring small anchors. While in [29], Ma et al. presented the rotation region proposal network (RRPN) to detect arbitrarily oriented text. This paper aims to generate tilted proposals with angular information about the text orientation. The angle information is then adjusted and bounding box regression is performed to make the proposals more accurately fit the orientation of the text region.

Previous research has adopted bounding boxes or quadrangles as a text description approach. For example, the approach presented in [3] was based on a single shot MultiBox detector (SSD) [30] object detection framework, which used a quadrilateral or rotated rectangle representation to replace the rectangular box. Reference [5] proposed an end-to-end two-stage scene text detection network architecture, named the quadrilateral region proposal network (QRPN), that can accurately locate scene texts with quadrilateral boundaries. In [31], the authors proposed the rotation-sensitive regression detector (RRD) framework to perform classification and regression on different features extracted by two different

designs of network branches. Deng et al. [32] proposed a new two-stage algorithm. In the first stage, the method predicts text instance locations by detecting and linking corners instead of traditional anchor points. In the second stage, the authors designed a pooling layer called dual-Roi pooling, which embeds data augmentation inside a regional sub-network.

Anchor-free text detector. Anchor-free-based approaches treat text as a distinct object and leverage efficient object detection architectures (e.g., YOLOv1 [33], SSD [30], CornerNet [34], and DenseBox [35]) to detect words or text lines directly from natural images. YOLOv1 [33] does not use anchor boxes, but rather predicts bounding boxes at points close to the center of the object, resulting in a low recall. CornerNet [34] is a recently introduced single-stage anchor-free detector that detects bounding box corner pairs and groups them together to make the final detected bounding box. However, CornerNet needs a complex postprocessing procedure to cluster corner pairs that belong to the same instance. An additional distance metric also needs to be learned when grouping. Another family of anchor-free detectors is that based on DenseBox [35] (e.g., UnitBox [36]). These detectors are considered unsuitable for generic object detection due to difficulties in handling overlapping bounding boxes and relatively low recall values. FCOS [22] is a single-stage anchor-free detector recently proposed to obtain detection accuracy comparable to traditional anchor-based detectors. Unlike YOLOv1, FCOS utilizes all points in the ground truth bounding box to predict the bounding box, while the detected low-quality bounding boxes are restrained by the proposed “centerness” branch. In this paper, we introduce a method based on FCOS, and integrate the Bidirectional Feature Pyramid Network (BiFPN) into the FCOS framework. Experiments demonstrate the ability of BiFPN to enhance the model learning capacity and increase the receptive field.

3. Our Approach

3.1. Bidirectional Feature Pyramid Network

Mainstream text detection architectures employ pyramid feature combination steps (e.g., feature pyramid network (FPN)) to enrich features with high-level semantic information. The traditional FPN generally enriches the feature maps from the final output of a single path architecture in a top-down manner. Despite their great success, such methods are limited by several factors: (1) the design does not incorporate high-level context with the former level features, retaining the spatial detail and semantic information in the network path. (2) Input features vary with resolution, resulting in inconsistent contributions to the output feature. Tan et al. [37] recently proposed a bidirectional pyramid network (BiFPN) that fused multiscale features for object detection. The framework contained two key modules: cross-scale connections and weighted feature fusion. Unlike the one-way information flow of the traditional top-down FPN, BiFPN included a bottom-up path aggregation network and an additional edge from the original input to the output node. We employed five levels of feature maps defined as {P3, P4, P5, P6, P7} (Figure 2), with each feature level P3, P4, P5, P6, and P7 having 8, 16, 32, 64, and 128 strides, respectively.

The fast feature fusion of the BiFPN, as follows:

$$\begin{aligned} P_3^{out} &= \text{Conv} \left(\frac{w_1 \cdot P_3^{in} + w_2 \cdot \text{Resize}(P_4^{in})}{w_1 + w_2 + \epsilon} \right), \\ P_4^{td} &= \text{Conv} \left(\frac{w_1 \cdot P_4^{in} + w_2 \cdot \text{Resize}(P_5^{in})}{w_1 + w_2 + \epsilon} \right), \\ P_4^{out} &= \text{Conv} \left(\frac{w_1 \cdot P_4^{in} + w_2 \cdot P_4^{td} + w_3 \cdot \text{Resize}(P_3^{out})}{w_1 + w_2 + w_3 + \epsilon} \right), \\ P_7^{out} &= \text{Conv} \left(\frac{w_1 \cdot P_7^{in} + w_2 \cdot \text{Resize}(P_6^{out})}{w_1 + w_2 + \epsilon} \right), \end{aligned} \quad (1)$$

where P_n^{td} is the middle result of the n th layer on the top-down path; P_n^{out} is the output result of the n th layer on the bottom-up path; Resize is an upsampling or downsampling operation for resolution matching; Conv is depthwise separable convolution [38] for feature

fusion, and here we use weighted normalized feature fusion [37]; $w_i \geq 0$ is guaranteed by applying a Relu after each w_i and $\epsilon = 0.0001$.

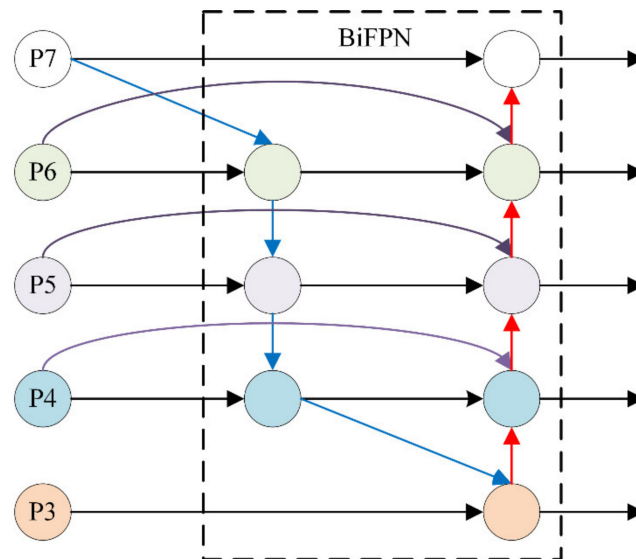


Figure 2. Architecture of proposed feature network.

3.2. FCOS for Text Detection

The majority of state-of-the-art text detectors such as Deep-text [1], TextBoxes [2], TextBoxes++ [3] and ABCNet [39] use a predefined anchor box, which requires elaborate parameter tuning and complex calculations for box IoUs during training. With no anchor box, the FCOS [22] can predict the 4D vector and class labels for each spatial location on the feature map layer directly. The 4D vector describes the relative offsets (l , t , r , and b) from the four sides (left, top, right, and bottom) of a location bounding box (Figure 3).



Figure 3. The proposed method predicts a 4D vector (l , t , r , b), where l , t , r , b are the distances from the location to the four sides of the bounding box.

Anchor-Free Text Detection Head. Let $F_i \in \mathbb{R}^{H \times W \times C}$ be the feature maps at layer i of a backbone CNN and s be the total stride until the layer. The ground-truth bounding boxes for an input image are defined as $\{B_i\}$, where $B_i = (x_0^{(i)}, y_0^{(i)}, x_1^{(i)}, y_1^{(i)}) \in \mathbb{R}^4$. Here $(x_0^{(i)}, y_0^{(i)})$ and $(x_1^{(i)}, y_1^{(i)})$ denote the coordinates of the left-top and right-bottom corners of the bounding box. We can map each location $p(x, y)$ on feature map F_i back onto the input image using $(\lfloor \frac{s}{2} \rfloor + xs, \lfloor \frac{s}{2} \rfloor + ys)$. This is close to the center of the receptive field

of location p . Anchor-based text detectors consider the location on the input image as the anchor box center and regress the target bounding box with these anchor boxes as references. In contrast, following [19], we treated the location as a training sample instead of an anchor-box and regressed the target bounding box at the location.

In our framework, p was treated as a positive sample if it fell into any ground-truth box. In addition to the classification label, we also defined the 4D real vector $t^* = (l^*, t^*, r^*, b^*)$ as the regression targets at the location, where l^*, t^*, r^*, b^* are the distances from the location to the four edges of the bounding box (as shown in Figure 3). We simply selected the bounding box with the minimal area as the regression target. More specifically, if location (x, y) was associated with bounding box B_i , the training regression targets for the location can be determined as Equation (2).

$$\begin{aligned} l^* &= x - x_0^{(i)}, t^* = y - y_0^{(i)} \\ r^* &= x_1^{(i)} - x, b^* = y_1^{(i)} - y \end{aligned} \quad (2)$$

After the execution of the feature extraction backbone, the anchor-free text detection head predicted the text location in the images of nature. Figure 4 presents the network architecture of the text box detection network. Similar to [40], the input features of the backbone network were fed into the three convolutional layers for the final text/nontext classification and the quadrilateral bounding box regression branches. Note that the proposed method has at least $9 \times$ fewer network output variables than popular anchor-based text detectors [1,3] that use preset anchor boxes. Following [22], we also employed the centerless branch to eliminate low-quality text prediction bounding boxes.

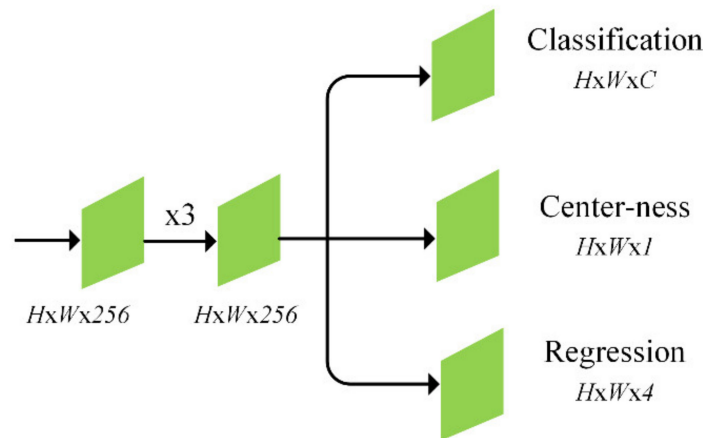


Figure 4. Architecture of text box prediction network, where H and W are the height and width of the feature maps.

4. Experiments

4.1. Datasets

We evaluated our proposed method on several standard benchmark tasks including ICDAR 2013 (IC13) [41] and ICDAR 2015 (IC15) [42] for multioriented text detection and ICDAR 2017 MLT (MLT17) [43] for multilingual text detection. IC13 [41] inherits from ICDAR 2003 [44], with 229 and 233 natural images for training and testing, respectively. IC15 [42], the first incidental scene text dataset, was built for the Incidental Scene Text challenge in the ICDAR-2015 Robust Reading Competition and contains 1000 images for training and 500 images for validation/testing. The 17,548 text instances (annotated by the 4 vertices of the quadrangle) are usually skewed or blurred since they are acquired without prior preference or intention. IC15 provides word-level English annotations. MLT17 [43] consists of 18,000 images containing text in 9 different languages: Arabic, Bengali, Chinese, English, French, German, Italian, Japanese, and Korean. A total of 9000 images were

used for training the model (7200 for training and 1800 for validation), and the other half for testing.

In order to compare with the state-of-the-art methods, we performed the comparison on three popular public datasets. Specifically, we used the official evaluation tools in the public datasets ICDAR 2013, ICDAR 2015, while for ICDAR 2017, we used the evaluation tools provided by the authors [45].

4.2. Implementation Details

We used EfficientNet-B1 [46] as the backbone networks for our proposed model, with the hyper-parameters following those of EfficientDet [37]. In particular, our network was trained using the stochastic gradient descent (SGD) across 80 K iterations with the initial learning rate of 0.01 and a minibatch of 16 images. The learning rate was reduced by a factor of 10 at iterations 50 K and 70 K. Furthermore, the weight decay and momentum were set as 0.0005 and 0.9, respectively. We pretrained the weights on ImageNet [47] for the initialization of our backbone networks, while the newly added layers were initialized by applying random weights with a gaussian distribution of mean 0 and standard deviation of 0.01. For the ICDAR-2017 MLT dataset, we used the training and validation data (i.e., 9000 training images), while for both ICDAR 2013 and ICDAR 2015, we employed the pretrained model from ICADAR-2017 MLT, with the provided training images applied for finetuning. We implemented our method in Py torch and performed the training on a RTX TITAN GPU system.

Loss Function: The loss function employed during training is defined as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{box} + \mathcal{L}_{center}, \quad (3)$$

where the classification loss $\mathcal{L}_{cls}$, box regression loss $\mathcal{L}_{box}$, and centerness loss $\mathcal{L}_{center}$ are equal to those in [22].

4.3. Results and Comparison

We compared the performance of our approach with that of the state-of-the-art methods using the ICDAR 2013 (IC13), ICDAR 2015 (IC5), and ICDAR 2017 MLT (MLT17) benchmark datasets (Tables 1–3).

Table 1. Our method and state-of-the-art methods of detection results on the IC13 dataset.

Work	P	R	F
FASText [48]	84	69	77
DeepText [1]	85	81	83
WeText [49]	82.6	93.6	87.7
TextBoxes [2]	89	83	86
R2CNN [26]	92	81	86
SegLink [50]	87.7	83	85.3
SLPR [51]	90	72	80
RGC [52]	89	77	83
Proposed + FPN	89.6	81.7	85.92
Proposed + BiFPN	93.1	84.6	88.65

Comparison with Prior Works. Tables 1–3 demonstrate that our approach outperformed the other methods for all three datasets with just the use of single-scale and single-model testing. For example, our approach achieved values of 93.1, 84.6, and 88.65 for the precision, recall, and F-measure, respectively, on the challenging ICDAR-2013 dataset (Table 1). Equivalent values for the ICDAR-2015 dataset were 87.6, 84.91, and 86.23, respectively, surpassing the other methods despite their use of extra training data (Table 2). The same trend was observed for the MLT17 dataset, with values of 83.41, 78.26, and 80.75, respectively (Table 3). Our proposed method achieved better results than the other methods on these three challenging text detection benchmarks. Figure 5 depicts the qualitative detection results.

Table 2. Arbitrary quadrilateral text detection results on the IC15 dataset.

Work	P	R	F
WordSup [53]	77.03	79.33	78.16
SSTD [54]	80	73	77
EAST [55]	83.27	78.3	80.72
FOTS [56]	88.8	82	85.3
TextBoxes++ [3]	87.8	78.5	82.9
[57]	94.1	70.7	80.7
Proposed + FPN	85.6	82.43	83.99
Proposed + BiFPN	87.6	84.91	86.23

Table 3. Multilingual text detection results on the MLT17 dataset. * denotes multiscale testing.

Work	P	R	F
FOTS [56]	80.95	57.51	67.25
FOTS * [56]	81.80	62.30	70.70
SPCNET [58]	80.60	68.60	74.10
[59]	88.60	73.90	80.50
[57]	83.80	55.60	72.40
PSENet [60]	77.01	68.40	72.45
Pixel-Anchor [61]	79.54	59.54	68.10
Pixel-Anchor * [61]	83.90	65.80	73.76
[21]	80.00	69.80	74.30
Proposed + FPN	80.12	76.42	78.23
Proposed + BiFPN	83.41	78.26	80.75

**Figure 5.** Detection results of the proposed FCOS based text detector.

Our framework had the following advantages.

- Scene text detection was designed as a proposal-free and anchor-free pipeline, which did not require the manual design of an anchor box or the heuristic adjustment of an anchor box, reducing the number of parameters and simplifying the training process.
- Compared to the anchor box-based methods, our one-stage text detector, which avoided the use of RPN networks and IOU-based proposal filtering, greatly reduced the computation.
- As a result, our text detection framework, was simpler and more efficient. Our framework could be easily extended to other vision tasks, providing a new solution for the detection task of scene text recognition.

BiFPN is better than FPN. We compared BiFPN with FPN on the three benchmark datasets. As shown in Tables 1–3, the integration of BiFPN into the proposed approach improved the F-measure by 2.7, 2.2 and 2.5% compared to the proposed approach with the traditional FPN. In our experiments, we observed that BiFPN had better feature extraction ability than FPN. The main reason for this was BiFPN’s two-way feature fusion, which could better preserve text features.

5. Conclusions

In the current paper, we proposed a new FCOS based text detection approach that includes an anchor-box and proposal-free one-stage text detector. Our method can robustly detect text from natural scene images, which is simpler, more efficient, and more scalable than anchor-based methods. Moreover, we demonstrated the ability of the bidirectional feature pyramid network (BiFPN) as the new backbone network of FCOS to significantly enhance the feature representation of FCOS, effectively improving text detection in natural scene pictures. Our proposed method achieved better results than other state-of-the-art methods on these three challenging text detection benchmarks (ICDAR 2013, ICDAR 2015 and ICDAR 2017 MLT). In the future, we will continue to focus on feature fusion methods to further improve detection capabilities. In addition, due to the simplicity of our framework, we are interested in extending our framework to scene text recognition tasks.

Author Contributions: Conceptualization, D.C., J.D. and Y.Z.; methodology, D.C., J.D. and Y.Z.; software, D.C., J.D. and Y.Z.; validation, D.C., J.D. and Y.Z.; formal analysis, D.C. and Y.Z.; resources, D.C., J.D. and Y.Z.; data curation, D.C., J.D. and Y.Z.; writing—original draft preparation, D.C., and J.D.; writing—review and editing, D.C. and Y.Z.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Science & Technology Department of Sichuan Province, grant number 2020ZHZY0002.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This research was supported by the Science & Technology Department of Sichuan Province (Grant No. 2020ZHZY0002). The authors sincerely thank Teddy Zhang who provided valuable comments in writing this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhong, Z.; Jin, L.; Zhang, S.; Feng, Z. DeepText: A Unified Framework for Text Proposal Generation and Text Detection in Natural Images. *arXiv* **2016**, arXiv:1605.07314.
2. Liao, M.; Shi, B.; Bai, X.; Wang, X.; Liu, W. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
3. Liao, M.; Shi, B.; Bai, X. TextBoxes++: A Single-Shot Oriented Scene Text Detector. *IEEE Trans. Image Process.* **2018**, *27*, 3676–3690. [[CrossRef](#)] [[PubMed](#)]

4. Liu, Y.; Jin, L. Deep Matching Prior Network: Toward Tighter Multi-oriented Text Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3454–3461.
5. Wang, S.; Liu, Y.; He, Z.; Wang, Y.; Tang, Z. A quadrilateral scene text detector with two-stage network architecture. *Pattern Recognit.* **2020**, *102*, 7230. [[CrossRef](#)]
6. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)] [[PubMed](#)]
7. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651.
8. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
9. Tian, Z.; He, T.; Shen, C.; Yan, Y. Decoders Matter for Semantic Segmentation: Data-Dependent Decoding Enables Flexible Feature Aggregation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3121–3130. [[CrossRef](#)]
10. Liu, Y.; Chen, K.; Liu, C.; Qin, Z.; Luo, Z.; Wang, J. Structured Knowledge Distillation for Semantic Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Institute of Electrical and Electronics Engineers (IEEE), Long Beach, CA, USA, 15–20 June 2019; pp. 3121–3130. [[CrossRef](#)]
11. Chen, Y.; Shen, C.; Wei, X.; Liu, L.; Yang, J. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), Venice, Italy, 22–29 October 2017; pp. 1221–1230.
12. Luo, C.; Chu, X.; Yuille, A. OriNet: A Fully Convolutional Network for 3D Human Pose Estimation. *arXiv* **2018**, arXiv:1811.04989.
13. Yin, W.; Liu, Y.; Shen, C.; Yan, Y. Enforcing Geometric Constraints of Virtual Normal for Depth Prediction. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 5683–5692. [[CrossRef](#)]
14. Liu, F.; Shen, C.; Lin, G.; Reid, I. Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 2024–2039. [[CrossRef](#)]
15. Zhang, Z.; Zhang, C.; Shen, W.; Yao, C.; Liu, W.; Bai, X. Multi-oriented Text Detection with Fully Convolutional Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Institute of Electrical and Electronics Engineers (IEEE), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4159–4167. [[CrossRef](#)]
16. He, D.; Yang, X.; Liang, C.; Zhou, Z.; Ororbia, A.G.; Kifer, D.; Giles, C.L. Multi-scale FCN with Cascaded Instance Aware Segmentation for Arbitrary Oriented Word Spotting in the Wild. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 22–25 July 2017; pp. 474–483. [[CrossRef](#)]
17. Du, C.; Wang, C.; Wang, Y.; Feng, Z.; Zhang, J. TextEdge: Multi-oriented Scene Text Detection via Region Segmentation and Edge Classification. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, 20–25 September 2019; pp. 375–380. [[CrossRef](#)]
18. Yao, C.; Bai, X.; Sang, N.; Zhou, X.; Zhou, S.; Cao, Z. Scene Text Detection via Holistic, Multi-Channel Prediction. *arXiv* **2016**, arXiv:1606.09002.
19. Bazazian, D. Fully Convolutional Networks for Text Understanding in Scene Images. *ELCVIA Electron. Lett. Comput. Vis. Image Anal.* **2020**, *18*, 6–10. [[CrossRef](#)]
20. Liao, M.; Wan, Z.; Yao, C.; Chen, K.; Bai, X. Real-Time Scene Text Detection with Differentiable Binarization. In Proceedings of the AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence (AAAI), New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11474–11481. [[CrossRef](#)]
21. Huang, Z.; Zhong, Z.; Sun, L.; Huo, Q. Mask R-CNN With Pyramid Attention Network for Scene Text Detection. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Hilton Waikoloa Village, HI, USA, January 7–11 January 2019; pp. 764–772. [[CrossRef](#)]
22. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Adelaide, Australia, 1 October 2019.
23. Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Reading Text in the Wild with Convolutional Neural Networks. *Int. J. Comput. Vis.* **2016**, *116*, 1–20. [[CrossRef](#)]
24. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
25. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
26. Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; Luo, Z. R2CNN: Rotational Region CNN for Orientation Robust Scene Text Detection. *arXiv* **2017**, arXiv:1706.09579.
27. Tian, Z.; Huang, W.; He, T.; He, P.; Qiao, Y. Detecting Text in Natural Image with Connectionist Text Proposal Network. *Math. Comput. Music* **2016**, *56*, 56–72. [[CrossRef](#)]
28. Xiang, D.; Guo, Q.; Xia, Y. Robust Text Detection with Vertically-Regressed Proposal Network. *Med. Image Comput. Comput. Assist. Interv.* **2016**, *2020*, 351–363. [[CrossRef](#)]

29. Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-Oriented Scene Text Detection via Rotation Proposals. *IEEE Trans. Multimed.* **2018**, *20*, 3111–3122. [\[CrossRef\]](#)
30. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. Machine Learning and Knowledge Discovery in Databases. *Appl. Data Sci. Demo Track* **2016**, 21–37. [\[CrossRef\]](#)
31. Liao, M.; Zhu, Z.; Shi, B.; Xia, G.-S.; Bai, X. Rotation-Sensitive Regression for Oriented Scene Text Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5909–5918. [\[CrossRef\]](#)
32. Deng, L.; Gong, Y.; Lin, Y.; Shuai, J.; Tu, X.; Zhang, Y.; Ma, Z.; Xie, M. Detecting multi-oriented text with corner-based region proposals. *Neurocomputing* **2019**, *334*, 134–142. [\[CrossRef\]](#)
33. Uijlings, J.R.R.; Van De Sande, K.E.A.; Gevers, T.; Smeulders, A.W.M. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [\[CrossRef\]](#)
34. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. *Int. J. Comput. Vis.* **2020**, *128*, 642–656. [\[CrossRef\]](#)
35. Huang, L.; Yang, Y.; Deng, Y.; Yu, Y. DenseBox: Unifying Landmark Localization with end to end Object Detection. *arXiv* **2015**, arXiv:1509.04874.
36. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T.S. UnitBox: An advanced object detection network. In Proceedings of the 2016 ACM on International Workshop on Security and Privacy Analytics, Washington, DC, USA, 15–19 October 2016; pp. 516–520. [\[CrossRef\]](#)
37. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787.
38. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
39. Liu, Y.; Chen, H.; Shen, C.; He, T.; Jin, L.; Wang, L. ABCNet: Real-Time Scene Text Spotting With Adaptive Bezier-Curve Network. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9806–9815. [\[CrossRef\]](#)
40. Lin, T.Y.; Goyal, P.; Girshick, R.B.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; I Bigorda, L.G.; Mestre, S.R.; Mas, J.; Mota, D.F.; Almazan, J.A.; Heras, L.P.D.L. ICDAR 2013 Robust Reading Competition. In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 1484–1493. [\[CrossRef\]](#)
42. Karatzas, D.; Bigorda, G.L.; Nicolaou, A.; Ghosh, S.K.; Bagdanov, A.D.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V.R.; Lu, S.; et al. ICDAR 2015 competition on Robust Reading. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Nancy, France, 23–26 August 2015; pp. 1156–1160. [\[CrossRef\]](#)
43. Nayef, N.; Yin, F.; Bizid, I.; Choi, H.; Feng, Y.; Karatzas, D.; Luo, Z.; Pal, U.; Rigaud, C.; Chazalon, J.; et al. ICDAR2017 Robust Reading Challenge on Multi-Lingual Scene Text Detection and Script Identification–RRC-MLT. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Institute of Electrical and Electronics Engineers (IEEE), Kyoto, Japan, 9–12 November 2017; Volume 1, pp. 1454–1459. [\[CrossRef\]](#)
44. Lucas, S.; Panaretos, A.; Sosa, L.; Tang, A.; Wong, S.; Young, R. ICDAR 2003 robust reading competitions. In Proceedings of the Seventh International Conference on Document Analysis and Recognition, Seoul, Korea, 29 August–1 September 2005; pp. 682–687. [\[CrossRef\]](#)
45. Yuliang, L.; Lianwen, J.; Shuaitao, Z.; Sheng, Z. Detecting Curve Text in the Wild: New Dataset and New Solution. *arXiv* **2017**, arXiv:1712.02170.
46. Tan, M.; Le, Q.V. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv* **2019**, arXiv:1905.11946.
47. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei, F.L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [\[CrossRef\]](#)
48. Buta, M.; Neumann, L.; Matas, J. FASText: Efficient Unconstrained Scene Text Detector. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015; pp. 1206–1214. [\[CrossRef\]](#)
49. Tian, S.; Lu, S.; Li, C. WeText: Scene Text Detection under Weak Supervision. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1501–1509. [\[CrossRef\]](#)
50. Shi, B.; Bai, X.; Belongie, S. Detecting Oriented Text in Natural Images by Linking Segments. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3482–3490. [\[CrossRef\]](#)
51. Zhu, Y.; Du, J. Sliding Line Point Regression for Shape Robust Scene Text Detection. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 3735–3740. [\[CrossRef\]](#)
52. Mohanty, S.; Dutta, T.; Gupta, H.P. Recurrent Global Convolutional Network for Scene Text Detection. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 2750–2754. [\[CrossRef\]](#)
53. Hu, H.; Zhang, C.; Luo, Y.; Wang, Y.; Han, J.; Ding, E. WordSup: Exploiting Word Annotations for Character Based Text Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4950–4959. [\[CrossRef\]](#)
54. He, P.; Huang, W.; He, T.; Zhu, Q.; Qiao, Y.; Li, X. Single Shot Text Detector with Regional Attention. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3066–3074. [\[CrossRef\]](#)

-
55. Zhou, X.; Yao, C.; Wen, H.; Wang, Y.; Zhou, S.; He, W.; Liang, J. EAST: An Efficient and Accurate Scene Text Detector. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Institute of Electrical and Electronics Engineers (IEEE), Honolulu, HI, USA, 21–26 July 2017; pp. 2642–2651. [\[CrossRef\]](#)
 56. Liu, X.; Liang, D.; Yan, S.; Chen, D.; Qiao, Y.; Yan, J. FOTS: Fast Oriented Text Spotting with a Unified Network. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5676–5685. [\[CrossRef\]](#)
 57. Lyu, P.; Yao, C.; Wu, W.; Yan, S.; Bai, X. Multi-oriented Scene Text Detection via Corner Localization and Region Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7553–7563. [\[CrossRef\]](#)
 58. Xie, E.; Zang, Y.; Shao, S.; Yu, G.; Yao, C.; Li, G. Scene Text Detection with Supervised Pyramid Context Network. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 9038–9045. [\[CrossRef\]](#)
 59. Dasgupta, K.; Das, S.; Bhattacharya, U. Scale-Invariant Multi-Oriented Text Detection in Wild Scene Image. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 2041–2045. [\[CrossRef\]](#)
 60. Wang, W.; Xie, E.; Li, X.; Hou, W.; Lu, T.; Yu, G.; Shao, S. Shape Robust Text Detection with Progressive Scale Expansion Network. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9328–9337. [\[CrossRef\]](#)
 61. Li, Y.; Yu, Y.; Li, Z.; Lin, Y.; Xu, M.; Li, J.; Zhou, X. Pixel-Anchor—A Fast Oriented Scene Text Detector with Combined Networks. *arXiv* **2018**, arXiv:1811.07432.