POLYGON-FREE: UNCONSTRAINED SCENE TEXT DETECTION WITH BOX ANNOTATIONS

Weijia Wu[1], *Enze Xie*[2], *Ruimao Zhang*[3], *Wenhai Wang*[4], *Ping Luo*[2], *Hong Zhou*[1]

[1] Zhejiang University
[2] The University of Hong Kong
[3] The Chinese University of Hong Kong, Shenzhen China
[4] Shanghai Artificial Intelligence Laboratory, China

ABSTRACT

Unlike existing works that employ fully-supervised training with polygon annotations, this study proposes an unconstrained text detection system termed Polygon-free (PF), in which most existing polygon-based text detectors (*e.g.,* PSENet [1]) are trained with only upright bounding box annotations. Our core idea is to transfer knowledge from synthetic data to real data to enhance the supervision information of upright bounding boxes. This is made possible with a simple segmentation network, namely Skeleton Attention Segmentation Network (SASN), that includes three vital components (*i.e.,* channel attention, spatial attention and skeleton attention map) and one soft cross-entropy loss.

Experiments demonstrate that the proposed Polygon-free yields surprisingly high-quality pixel-level results with only upright bounding box annotations. For example, without using polygon annotations, PSENet achieves an 80.5% F-score on TotalText (vs. 80.9% of fully supervised counterpart), **31.1%** better than training directly with upright bounding box annotations, and saves **80%+** labeling costs.

***Index Terms*—** Text Detection, Weakly Supervision

1. INTRODUCTION

Recently, most scene text detectors [1, 2] have utilized polygon annotation with many coordinates (see the example in Fig. 2 (d)) to capture texts with different shapes. Although the polygon annotations are more accurate than the upright bounding box annotations, the labeling cost of polygons is extremely high, limiting the wide use in large-scale real-world applications. By contrast, the upright bounding box annotations are more economical, and are 4× cheaper than polygons annotations [3, 4], *e.g.,* saving 80% + annotation cost on TotalText [3], as shown in Fig. 1. This cost gap becomes larger for the large-scale benchmarks such as ICDAR2019-LSVT [5] and ICDAR2019-Art [6].

To address the text data cost issue, we first propose a simple, unconstrained system termed Polygon-free (PF) for training text detector with upright bounding box annotated data.

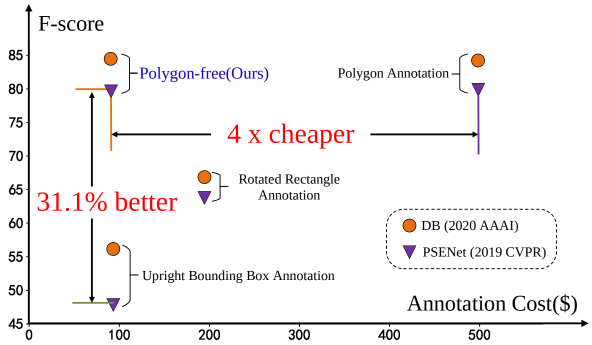
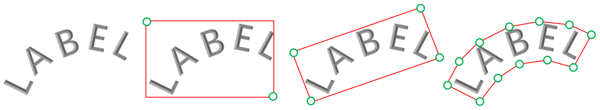


Fig. 1: **The performance and annotation cost for PSENet [1] and DB [7] on TotalText [3].** PSENet with Polygon-free is 31.1% better than training directly with upright bounding box annotation, 4× times cheaper than that with polygon annotation.

Compared with polygon annotations, the upright bounding box is much less expensive but contains less pixel information for efficient supervision. Thus, effectively utilizing the upright bounding box annotations and boosting the text detection performance becomes critical in this case. An alternative approach is to utilize synthetic text data [8, 9] that are largely available from the virtual world, and the ground truth can be freely and automatically generated. However, many previous works [8] have shown that training directly with synthetic data degrades the performance on real data due to a phenomenon known as "domain shift" (*e.g.,* 58.0% for EAST directly training on SynthText and testing on ICDAR2015). Unlike the existing works, motivated by the attention mechanism studies [10] and BoxSup [11], we propose a Skeleton Attention Segmentation Network (SASN), and carefully design a Skeleton Attention Module based on channel attention and spatial attention to reduce the domain shift, help the network learn domain-invariant features with stronger representation power for text in a prior upright bounding box. Besides, considering the particular geometry of texts, we argue that it is more important to focus on the skeleton than on other regions. Therefore, we introduce a soft attention weight map



(a) Text ($ 0) (b) Upright Bounding Box($ 1.5) (c) Rotated Rectangle ($ 3.0) (d) Polygon ($ 9.0)

Fig. 2: **Comparisons of annotation cost**. Annotation of more points is more expensive (the price of 100 text instances). Note that the data cost information is obtained from Amazon Mechanical Turk (MTurk).

called skeleton map and corresponding soft cross-entropy loss to enhance the representation power of the text skeleton.

To make it suitable for all detectors, the whole Polygon-free is divided into two steps: (1) We firstly train SASN with almost free synthetic data based on character annotation. And then, the box-level annotations are utilized to crop the real image, which are fed into the SASN for generating polygon-liked pseudo labels. (2) By splicing all of the local pseudo labels, the global pseudo label is obtained. In this way, upright bounding box annotations can be converted to high-quality polygon pseudo labels. General detectors (*e.g.,* PSENet [1]) trained on these pseudo labels can achieve almost the same performance as those trained on original polygon annotations. The main contributions are two folds:

(1) We demonstrate a simple, unconstrained Polygon-free system that can train most existing text detectors with only upright bounding box annotations. This means that general detectors (*e.g.,* PSENet [1]) can be trained by upright bounding box with no modification to the network itself.

(2) We introduce a Skeleton Attention Segmentation Network composed of three components (*i.e.,* Spatial, Channel, and Skeleton Attention) and a soft cross-entropy loss, which bridging the domain gap between synthetic data and real data.

2. RELATED WORK

Supervised Text Detection. Scene text detection has achieved remarkable progress in the deep learning era. Previous methods [12, 13] focused on horizontal or multi-oriented text detection. CTPN [14] adopted Faster RCNN [15] and modified RPN to detect texts. EAST [16] used FCN [17] to predict the text score map, distance map and angle map. Recent works focused on curved text detection [18, 19, 20]. PSENet [1] and PAN [2] treat text instances as kernels with different scales and reconstruct text instance with post-processing. There are a few previous works [21] concerning weakly supervised text detection. WordSup[22] trains a character detector by exploiting word annotations in rich large-scale real scene text datasets.

Attention Mechanism. One important property of a human visual system is that we know "what" and "where" to focus our attention in an image. Similarly, attention enables the artificial model to focus only on the important data. The classic works based on attention, BAM [10] and CBAM [10] increase the accuracy of the classifier by utilizing both 1D channel and 2D spatial self-attention maps.

3. APPROACH

3.1. Skeleton Attention Segmentation Network

Network Architecture. As presented in Fig. 3 (a), we use ResNet50 [23] as the backbone network for the SASN, and extract three levels of features (*i.e.,* $C1, C2, C3$) from different downsampled scales (*i.e.,* $1/4, 1/8, 1/16$). After that, the skeleton stream fuses C1 and C3 to predict the skeleton attention map. The skeleton attention map is downsampled to make it suitable for multi-scale feature maps. At the same time, the regular stream refines multi-scale features (*i.e.,* $C1, C2, C3$) by applying the Skeleton Attention Module. Finally, $C1, C2, C3$ are fed into the Decoder, as shown in Fig. 3 (c). The C3 and C2 are first fused by up-sampling and concatenation. The fused feature map is further up-sampled to fuse with $C1$ in the same approach.

3.2. Skeleton Attention Module

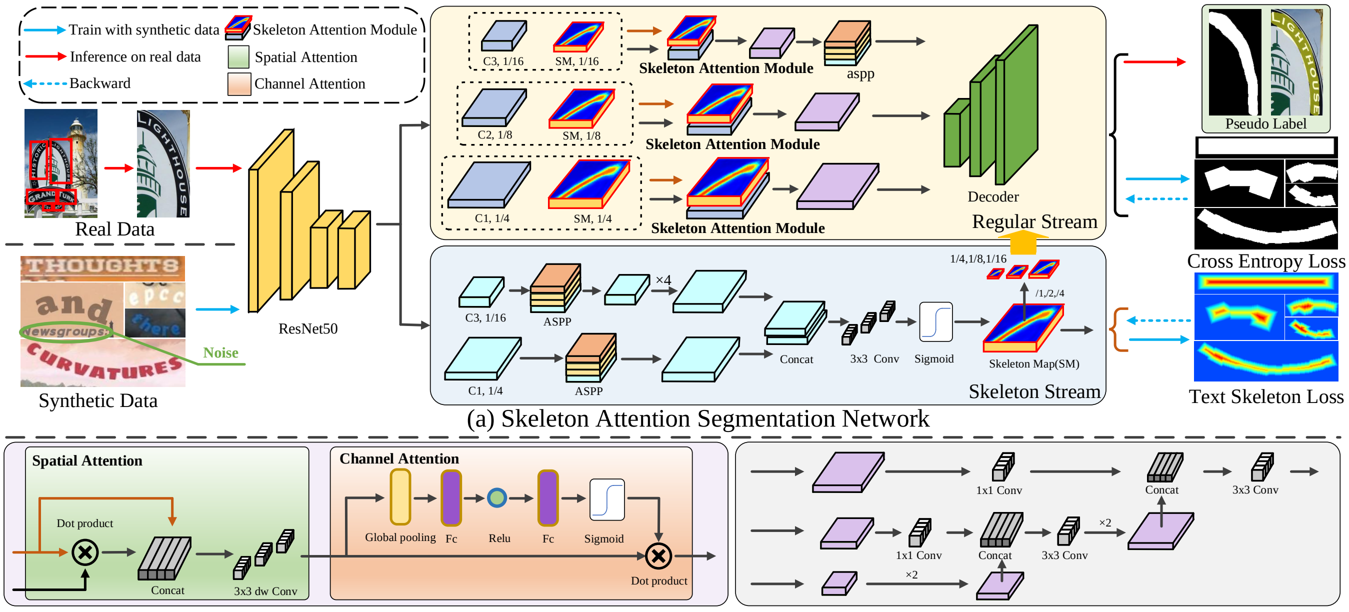
Spatial and Channel Attention. Fig. 3 (b) illustrates the details of the two attention methods. Given an intermediate feature map $F \in \mathbb{R}^{C\times H\times W}$ (*i.e.,* the orange input in Fig. 3 (b)) and an skeleton map $F_{sm} \in \mathbb{R}^{C\times H\times W}$ (*i.e.,* the black input in Fig. 3 (b)), spatial attention is first utilized to focus on "where" is the text skeleton by multiplication and concatenating. The spatial attention is computed as:

$$F^{'} = f^{3\times3}([F \otimes F_{sm}; F]) , \tag{1}$$

where $\otimes$ denotes element-wise multiplication, and $f^{3\times3}$ represents a convolution operation with the filter size of 3×3. Following the CBAM [10], the channel attention is utilized to focus on "what" is meaningful given an input image by exploiting the inter-channel relationship of the feature. Global pooling is first used to aggregate spatial information of the feature map $F^{'}$. Then, the fully connected layer is used to connect different channels, as each channel of a feature map is considered as a feature detector [24]. In short, a 1D channel attention map $M_c \in \mathbb{R}^{C\times1\times1}$ can be obtained by feeding the feature map $F^{'}$ into the $GlobalPool\text{-}Fc\text{-}Relu\text{-}Fc\text{-}Sigmoid$ layers. The final refined output from SAM can be computed as: $F^{''} = M_c \otimes F^{'}$.

In practice, the predicted skeleton map F_{sm} is downsampled to obtain multi-scale maps (*i.e.,* 1/4, 1/8, 1/16), that are transferred to the regular stream (the thick yellow arrow in Fig. 3 (a)). And then, the extracted feature map (*e.g.,* $C1$) and the corresponding scale skeleton map are used as the input of the Skeleton Attention Module. Note that the proposed Skeleton Attention Module is shared for the high-level and the low-level feature maps (*i.e.,* $C1, C2, C3$).

Skeleton Attention Map and Loss. Given an input sample $(x_i, y_i) \in \{(x_1, y_1), (x_2, y_2), ...(x_n, y_n)\}$, where x_i and



(a) Skeleton Attention Segmentation Network

(b) Skeleton Attention Module (SA)

(c) Decoder

Fig. 3: **The network architecture of Skeleton Attention Segmentation Network.** (a) SASN is composed of two streams: regular stream and skeleton stream. (b) Skeleton Attention Module is composed of channel attention and spatial attention, which refine the input feature map by weighting with text skeleton map. (c) The detailed structure of the Decoder.

y_i denote the i-th image and its labels. We use x_i^k to denote the k-th pixel of the i-th training image, with y_i^k= 0 for the background and y_i^k = 1 for the text pixel. To learn stronger representation of text skeleton, we define the text skeleton ground-truth as a soft label. For the k-th pixel in the text region of i-th image, we first calculate the shortest distance d_i^k between the k-th pixel to its nearest background pixel, and then the value p_i^k is defined as the soft skeleton label of k-th pixel by normalizing d_i^k to $[0, 1]$:

$$p_i^k = \frac{d_i^k}{d_i^*}, \tag{2}$$

where d_i^* is the maximum value of $\{d_i^k\}$ in the i-th image.

Intuitively, the pixels close to the skeleton of the text instance should have a greater value than the boundary pixels (see text skeleton label in Fig. 3 (a)). Since the soft label is a decimal representing the degree of distance, it is incompatible with the commonly binary cross-entropy loss. Besides, the L1 and L2 losses are not sensitive to the distance distribution among $[0, 1]$ [15]. Therefore, to handle the soft label, we modify the cross-entropy loss into a "soft" form. For a pixel x_i^k, p_i^k denotes the value of the k-th pixel in the ground truth skeleton map. $\mathcal{F}$ indicates networks. The soft cross-entropy loss for text skeleton loss in Fig. 3 (a) is defined as follows:

$$\mathcal{L}_{ske} = -\sum_k log(1 - \left|p_i^k - \mathcal{F}(x_i^k)\right|), \tag{3}$$

The whole loss function $\mathcal{L}$ for SASN can be expressed as a weighted sum of the loss for the regular stream $\mathcal{L}_{ce}$ (*i.e.,* the commonly binary cross-entropy loss) and the loss for the skeleton stream $\mathcal{L}_{ske}$:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda\mathcal{L}_{ske}, . \tag{4}$$

where λ is set to 2, which balances the importance between $\mathcal{L}_{ske}$ and $\mathcal{L}_{ce}$.

4. EXPERIMENTS

4.1. Ablation Study

Combining methods of Skeleton Attention. Tab. 2 shows the impact of three submodules: channel, spatial, and skeleton attentions. It is clear that the improvement due to channel attention is limited (*i.e.,* $+0.1\%$), but its spatial attention counterpart contributes to better gain (*i.e.,* $+0.7\%$), the main reason for this may be that spatial information is more important than semantic information for text segmentation task. Another important contribution is the soft attention weight map (*i.e.,* Skeleton Map), the performance achieves further improvement (*i.e.,* $+1.8\%$) after using the Skeleton Map. This is in line with our expectations, and we argue that the text skeleton is vital because of its high representation power.

4.2. Experiments on Scene Text Detection

Tab. 1 lists the results on the ICDAR2015, ICDAR2017 and MSRA-TD500 datasets. For *ICDAR2015* [26], PSENet with pseudo label achieves **almost the same performance** (85.5% v.s. 85.2%) with that using ground truth, proving the high quality of the pseudo label generated by PF. By contrast,

Method	Annotation	Pre	ICDAR2015/%			MSRA-TD500/%			ICDAR2017-MLT/%			Total-Text/%		
			P	R	F	P	R	F	P	R	F	P	R	F
Strong Supervision														
CTPN[14]	GT	-	74.2	51.6	60.9	-	-	-	-	-	-	-	-	-
EAST[16]	GT	-	80.5	72.8	76.4	81.7	61.6	70.2	-	-	-	-	-	-
PixelLink[25]	GT	-	82.9	81.7	82.3	83.0	73.2	77.8	-	-	-	-	-	-
PSENet[1]	GT	-	81.5	79.7	80.6	-	-	-	73.7	68.2	70.8	81.8	75.1	78.3
PSENet[1]	GT	✓	86.9	84.5	85.7	-	-	-	-	-	-	84.0	78.0	80.9
EAST†	GT	✓	82.0	82.4	82.2	77.9	76.5	77.2	70.3	62.8	66.4	-	-	-
PSENet†	GT	✓	86.4	84.0	85.2	84.1	85.0	84.5	72.5	69.1	70.8	83.4	78.1	80.7
Weakly Supervision														
EAST	b-GT	✓	70.8	72.0	71.4	48.3	42.4	45.2	67.2	60.1	63.5	-	-	-
EAST+PF	PL	✓	**81.3**	**82.2**	**81.8**	**77.4**	**75.5**	**76.4**	**67.6**	**64.9**	**66.3**	-	-	-
PSENet	b-GT	✓	72.7	74.3	73.5	47.5	39.5	43.1	66.4	63.1	64.7	51.9	47.5	49.6
PSENet+PF	PL	✓	**86.8**	**84.2**	**85.5**	**84.4**	**84.7**	**84.5**	**73.8**	**67.7**	**70.6**	**82.6**	**78.4**	**80.5**

Table 1: **The results of Polygon-free on ICDAR2015 [26], MSRA-TD500 [27], ICDAR2017-MLT [28], Total-Text [3].** † refers to our testing performance. The 'GT', 'b-GT' and 'PL' refer to the 'Ground Truth', 'Upright Bounding Box Ground Truth' and 'Pseudo Label from SASN', respectively. 'P', 'R', 'F' and "Pre" refer to 'Precision', 'Recall', 'F-score' and 'pretraining on SynthText', respectively. In green (strong supervision) and in **bold** (Polygon-free) are highlighted for comparison.

Method	Evaluation on Total-Text/%		
	Precision	Recall	F-score
BL	80.5	73.2	76.7
BL+SA(Cha)	79.8	74.0	76.8(+0.1)
BL+SA(Cha&Spa)	80.3	74.6	77.4(+0.7)
BL+SA(Cha&Spa&Skeleton Map)	**81.7**	**75.6**	**78.5**(+1.8)

Table 2: **Combining methods of Skeleton Attention**. 'SA', 'BL', 'Cha' and 'Spa' refer to 'Skeleton Attention', 'Baseline', 'Channel' and 'Spatial'.

direct training the detector with upright bounding box obtains an unsatisfactory F-score (73.5%), with a performance gap of more than 10%. For *MSRA-TD500* [27], annotations are provided at the line level, including the spaces between the words in the box. Therefore, bounding box on MSRA-TD500 usually contains a large background, causing poor performance (45.2% for EAST and 43.1% for PSENet). In this case, PSENet with pseudo label (84.5%) achieves a huge improvement (**+41.4**%) compared to training directly with upright bounding box (43.1%). For *Total-Text* [3], Tab. 1 lists the experimental results. The annotation on Total-Text is complex and polygonal in shape. The **great performance** (80.5%) of Polygon-free further demonstrates the significance of our work, and Fig. 4 provides the visualization of ground truth and pseudo label. Similar to MSRA-TD500, the upright bounding box on Total-Text also contains plenty of backgrounds, causing poor performance (45.0% and 49.6%). Using pseudo label generated by SASN can still achieve excellent performance (78.5% and 80.5%) with improvements of **33.5%** and **30.9%**.

Weakly supervision *v.s.* Strong supervision. To further present the effectiveness of Polygon-free, we summarized the results concerning three detectors (*i.e.,* PSENet [1], EAST [16] and DB [7]) and seven datasets (*i.e.,* ICDARs, Total-Text) to Table 3. As a **plug-and-play** weakly-supervised approach, the F1 of PF achieves 6.6% ∼ 35.5% improvements than directly training with upright bounding box (two points), almost equal to strong-supervised methods. This means that the proposed method can be directly applied for industry application with a little loss (*i.e.,* $< 0.5\%$) of performance when no need for any modification. The competitive performance of PF proves the practicability efficiency of the pseudo label. Fig. 4 gives some visualization of the pseudo label.

Datasets	Cited	Methods	F1▲/%	F1★/%	F1†/%
ICDAR2015	624	EAST	64.8	78.0 (+13.2)	76.4
		PSENet	73.5	85.5 (+12.0)	85.7
MSRA-TD500	630	EAST	35.2	70.7 (+35.5)	70.2
ICDAR2017MLT	128	PSENet	64.2	70.8 (+6.6)	70.8
Total-Text	142	PSENet	49.6	80.5 (+30.9)	80.9
		DB	49.1	84.5(+35.4)	84.7
CTW1500	105	PSENet	47.6	81.5(+33.9)	82.2
ICDAR2019LSVT	6	PSENet	57.3	77.6(+20.3)	77.4
ICDAR2019ArT	13	PSENet	46.6	69.2(+22.6)	69.5

Table 3: **Weakly supervision *v.s.* Strong supervision**. ▲, ★ and † refer to 'Upright Bounding Box', 'Polygon-free(ours)' and 'Original Paper Report or our testing result'.

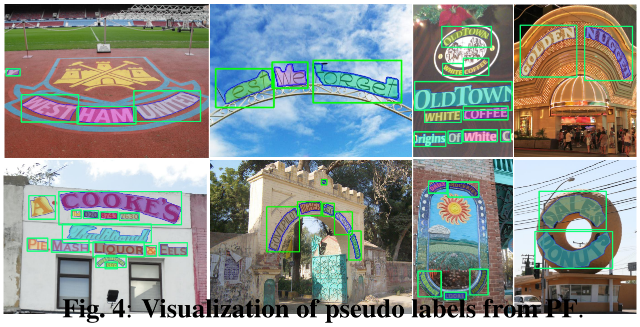


Fig. 4: **Visualization of pseudo labels from PF**.

5. CONCLUSION

In this paper, we present a simple but effective system termed Polygon-free, in which most existing polygon-based text detectors are trained with upright bounding box. The core is to transfer knowledge from synthetic to real data to enhance the supervision information via a skeleton attention segmentation network. The experiments showed that our method achieves almost the same performance as that of strong supervision while saving huge annotation cost, which can provide a new perspective for weakly supervised text detection.

6. REFERENCES

[1] Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao, "Shape robust text detection with progressive scale expansion network," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019, pp. 9336–9345. 1, 2, 4

[2] Wenhai Wang, Enze Xie, Xiaoge Song, Yuhang Zang, Wenjia Wang, Tong Lu, Gang Yu, and Chunhua Shen, "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2019, pp. 8440–8449. 1, 2

[3] Chee Kheng Ch'ng and Chee Seng Chan, "Total-text: A comprehensive dataset for scene text detection and recognition," in *Proc. Int. Conf. Document Analysis Recogn.*, 2017, pp. 935–942. 1, 4

[4] Liu Yuliang, Jin Lianwen, Zhang Shuaitao, and Zhang Sheng, "Detecting curve text in the wild: New dataset and new solution," *arXiv preprint arXiv:1712.02170*, 2017. 1

[5] Yipeng Sun, Jiaming Liu, Wei Liu, Junyu Han, Errui Ding, and Jingtuo Liu, "Chinese street view text: Large-scale chinese text reading with partially supervised learning," in *ICCV*, 2019, pp. 9086–9095. 1

[6] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al., "Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art," in *ICDAR2019*. IEEE, 2019, pp. 1571–1576. 1

[7] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai, "Real-time scene text detection with differentiable binarization," in *AAAI*, 2020, pp. 11474–11481. 1, 4

[8] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman, "Synthetic data for text localisation in natural images," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016, pp. 2315–2324. 1

[9] Shangbang Long and Cong Yao, "Unrealtext: Synthesizing realistic scene text images from the unreal world," *arXiv preprint arXiv:2003.10608*, 2020. 1

[10] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *eccv*, 2018, pp. 3–19. 1, 2

[11] Jifeng Dai, Kaiming He, and Jian Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *ICCV*, 2015, pp. 1635–1643. 1

[12] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu, "Textboxes: A fast text detector with a single deep neural network," in *Proc. AAAI Conf. Artificial Intell.*, 2017, pp. 4161–4167. 2

[13] Zheng Zhang, Chengquan Zhang, Wei Shen, Cong Yao, Wenyu Liu, and Xiang Bai, "Multi-oriented text detection with fully convolutional networks," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016, pp. 4159–4167. 2

[14] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proc. Eur. Conf. Comp. Vis.*, 2016, pp. 56–72. 2, 4

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proc. Advances in Neural Inf. Process. Syst.*, 2015, pp. 91–99. 2, 3

[16] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang, "East: an efficient and accurate scene text detector," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017, pp. 5551–5560. 2, 4

[17] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015, pp. 3431–3440. 2

[18] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee, "Character region awareness for text detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9365–9374. 2

[19] Yongchao Xu, Yukang Wang, Wei Zhou, Yongpan Wang, Zhibo Yang, and Xiang Bai, "Textfield: Learning a deep direction field for irregular scene text detection," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5566–5579, 2019. 2

[20] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai, "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," in *Proc. Eur. Conf. Comp. Vis.*, 2018, pp. 67–83. 2

[21] Shangxuan Tian, Shijian Lu, and Chongshou Li, "Wetext: Scene text detection under weak supervision," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1492–1500. 2

[22] Han Hu, Chengquan Zhang, Yuxuan Luo, Yuzhuo Wang, Junyu Han, and Errui Ding, "Wordsup: Exploiting word annotations for character based text detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4940–4949. 2

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 2

[24] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *eccv*. Springer, 2014, pp. 818–833. 2

[25] Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai, "Pixellink: Detecting scene text via instance segmentation," in *Proc. AAAI Conf. Artificial Intell.*, 2018, pp. 6773–6780. 4

[26] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Bagdanov, et al., "Icdar 2015 competition on robust reading," in *Proc. Int. Conf. Document Analysis Recogn.*, 2015, pp. 1156–1160. 3, 4

[27] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2012, pp. 1083–1090. 4

[28] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al., "Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt," in *ICDAR2017*. IEEE, 2017. 4