symmetry

MDPI

Review

Scene Text Detection in Natural Images: A Review

Dongping Cao [1,2], Yong Zhong [1,2,*], Lishun Wang [1,2], Yilong He [1,2] and Jiachen Dang [1,2]

[1] Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu 610041, China; caodongping16@mails.ucas.edu.cn (D.C.); wanglishun17@mails.ucas.edu.cn (L.W.); heyilong17@mails.ucas.edu.cn (Y.H.); dangjiachen18@mails.ucas.edu.cn (J.D.)

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

* Correspondence: zhongyong@casit.com.cn

Received: 20 October 2020; Accepted: 20 November 2020; Published: 26 November 2020

Abstract: Scene text detection is attracting more and more attention and has become an important topic in machine vision research. With the development of mobile IoT (Internet of things) and deep learning technology, text detection research has made significant progress. This survey aims to summarize and analyze the main challenges and significant progress in scene text detection research. In this paper, we first introduce the history and progress of scene text detection and classify the traditional methods and deep learning-based methods in detail, pointing out the corresponding key issues and techniques. Then, we introduce commonly used benchmark datasets and evaluation protocols and identify state-of-the-art algorithms by comparison. Finally, we summarize and predict potential future research directions.

Keywords: scene text detection; natural images; deep learning; detection framework

1. Introduction

Scene text detection (STD) is the process of detecting the presence and position of text in scene images. STD not only acts as a detection and positioning tool but also plays a key role in extracting important high-level semantic information from scene images. It has important applications in intelligent transportation systems [1], content-based image retrieval [2], industrial automation [3], portable vision systems [4,5], etc.

Evolution of STD. The concept of "scene text detection" first appeared in the 1990s [6–8]. With the rapid development of Internet technology and portable mobile devices, more and more scenarios have emerged where a need exists for extracting text from image information. At present, scene text detection has become a significant aspect of computer vision and pattern recognition techniques, as well as a research hotspot in the field of document analysis and recognition. Some top international conferences, such as ICDAR (International Conference on Document Analysis and Recognition), ICCV (International Conference on Computer Vision), ECCV (European Conference on Computer Vision), AAAI (AAAI Conference on Artificial Intelligence), include scene text detection listed as a separate research topic.

Motivation for writing this review. In 1998, Lecun et al. designed the LeNet5 model [9], which achieved a 99.1% recognition rate on the MNIST dataset. In recent years, deep learning (DL) has attracted significant attention due to its success in various domains, and DL-based STD methods with minimal feature engineering have been flourishing. A considerable number of studies have applied deep learning to STD and successively advanced the state-of-the-art performance [10–17]. This trend motivates us to conduct a review to report the current status STD technique research.

Contributions of this review. We thoroughly review the technological development of STD in order to inspire and guide researchers and practitioners in the field. Specifically, in Sections 3 and 4,

we categorize STD by technique into traditional detection methods and deep learning-based detection methods and representative techniques from both categories. Then, in Sections 5 and 6, we provide a comprehensive survey of STD benchmark datasets and evaluation methods. In addition, we summarize and analyze the most representative approaches to DL techniques for STD in Section 6.2. Finally, we introduce the challenges of STD and outline future directions for the field.

2. Background

In this section, we first introduce the definition of the STD and then summarize the important features of scene text in natural images.

2.1. What Is STD?

Scene text detection is to locate text in complex scene images. Examples of STD include text detection in various contexts, such as books, ID cards, tickets, intelligent traffic scenarios, such as road signs, license plate recognition, etc. Figure 1 gives a formal overview of the text detection process. We have summarized a unified three-stage framework that most existing models fit into. The three stages as follows:

1. **Transformation.** In this stage, the input image is transformed into a new one using a spatial transformation network (STN) [18] framework, while any text contained in it is rectified. The rectification process facilitated subsequent stages. It is powered by a flexible thin-plate spline (TPS) transformation, which can handle a variety of text irregularities [19] and diverse aspect ratios of text lines [20].
2. **Feature Extraction.** This stage maps the input image to a representation that focuses on the attributes relevant for character recognition while suppressing irrelevant features, such as font, color, size, and background. With convolutional networks entering a phase of rapid development after AlexNet's [21] success at the 2012 ImageNet competition, Visual Geometry Group Network (VGGNet), GoogleNet [22], RestNet [23], and DetNet [24] are often used as a feature extractor.
3. **Prediction.** Predicts the position of the text in the image, usually expressed as a coordinate point.

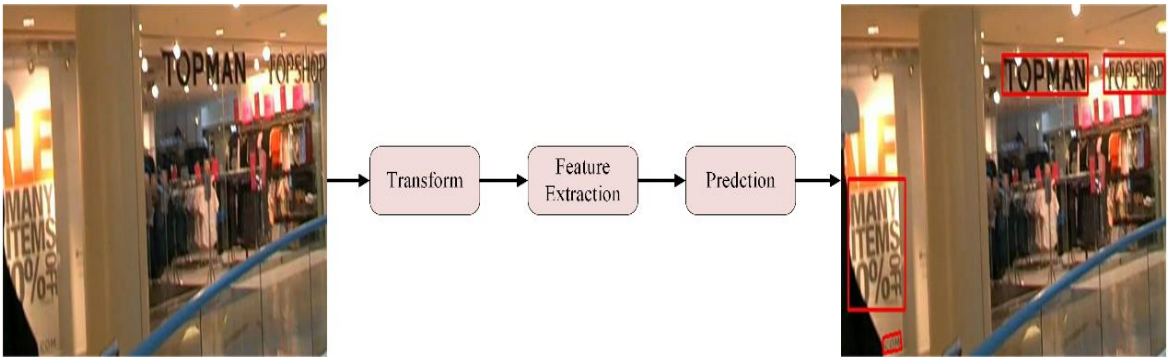

Figure 1. Visualization of an example flow of scene text detection. The image sample is from ICDAR2015 [25].

2.2. Features of Scene Text

Text detection in natural scene images is much more difficult than text detection in scanned document images because of the diversity of the forms the text may occur in. The main features of the scene text are summarized below:

1. Multiple languages may be mixed.
2. Characters may occur in different sizes, fonts, colors, brightness, contrast, etc.
3. Text lines may be horizontal, vertical, curved, rotated, twisted, or in other patterns.

4. The text area in the image may also be distorted (perspective, affine transformation), suffer defects, blurring, or other phenomena.
5. The background of scene images is extremely diverse, and text may appear on a plane, surface, or folded surface; the text region may be near complex interference textures, or non-text areas may have textures that approximate text, such as sand, grass, fences, brick walls, etc.

The rest of this paper is organized as follows. In Section 3, we take a look at the convolutional methods before the DL era. In Section 4, we introduce DL-based algorithms. In Sections 5 and 6, we describe the widely used benchmark datasets and evaluation protocols. Finally, we list the challenges and prospects of STD tasks.

3. Traditional Methods for STD

Traditional scene text detection methods follow two main technical routes; a sliding detection window (SDW)-based approach or a connected component analysis (CCA)-based approach. Those methods first acquire text candidate regions and then use handcrafted features to validate the candidate regions, and finally to obtain the text location information.

3.1. SDW Methods

Sliding detection window (SDW)-based methods use a top-down strategy to detect text. These methods [26–29] scan the entire scene image using a sliding window, extract the candidate text regions, and use a pre-trained classifier to identify whether the text is contained within the sub-window. Those classified as positive are further grouped into text regions using morphological operations [30,31], Support vector machines (SVMs) [32], Random Forests (RFs) [33,34], and artificial neural processors (ANPs) [35]. To cope with variable text sizes and line lengths effectively, this class of methods uses a multi-scale sliding window to obtain the text candidate area. In [36], a boosting framework integrating feature and weak classifier selection was employed to build an efficient text detector. In [31], a text detection method was proposed, which extracted six different feature classes and used AdaBoost with a multi-scale sequential search. SDW does not have to extract geometric features, such as edges and corner points of text from scene images, to obtain candidate text regions. Generally speaking, in circumstances where the text scale is small, or the contrast is weak, this type of approach is advantageous. However, it does not work well for skewed and curved text lines, and stride selection for the sliding windows is also an issue.

3.2. CCA Methods

CCA-based methods use a bottom-up strategy to detect text, which first extracts candidate components of similar properties (such as corner points [37], text texture [38,39], text boundary [40], and font color [41,42]) from the image and then uses manually formulated rules or automatically trained classifiers to filter out non-textual components [43].

Text in natural scene images is rich with edges and corner-point information. Therefore, some methods detect these elements to obtain text candidate regions and then classify these regions using rules or classifiers. In [44], a method based on edge-features for horizontally aligned artificial text detection from video images was proposed. In [28], edge detection was applied to obtain four edge maps, which represent the texture property of text, and then the k-means algorithm was used to detect the initial text candidates. In [45–47], some edge detection operators (e.g., Sobel, Canny) were used to detect the edge information of the image, and then morphological processing of the edge image was performed to eliminate pseudo-text regions. In [46], the FASText detector was proposed, which is based on an efficient pixel intensity comparison to surrounding pixels.

The most representative methods of this approach are the stroke width transform (SWT) [48] and the maximum stable extremal region (MSER) [49]. The SWT was first proposed in [48] (Figure 2) in 2010 and takes advantage based on the assumption that strokes located in the same text area have

approximately equal widths to obtain candidate text regions. The Canny algorithm was used to detect the edges of the input image, then the gradient direction of the edge pixels was calculated, and the algorithm searched for matching pixels along the path of the gradient direction. The sum of all pixels on the search path between matching pixels p and q was taken as the stroke width w between the two pixels. The method is simple, local, and data-dependent (see Figure 3), which makes it powerful enough to detect text in multiple fonts and languages. In [50], a method based on feature vectors of connected components generated through STW was proposed. For the formulation of the feature vectors, some properties were used, such as the directionality of the text edge gradients, high contrast with the background, and the geometry of the text components, jointly with the attributes found by the stroke width transformation. In [51], text and non-text regions were analyzed on three levels: pixel, component, and text-line levels. The stroke feature transform (SFT), which extends the SWT, was used as a low-level filter to determine whether a pixel belongs to text or not. In [52], an algorithm was proposed which used multiple low-level visual features to learn a model, which eventually provided a text attention map indicating candidate regions of text in the image. During detection, the text detector using SWT focused only on these selected image regions to reduce computation time and improve detection performance. In [53], an algorithm based on SWT was used to extract arbitrary text in natural scene images. References [54,55] involved a modified SWT for detecting scene text.

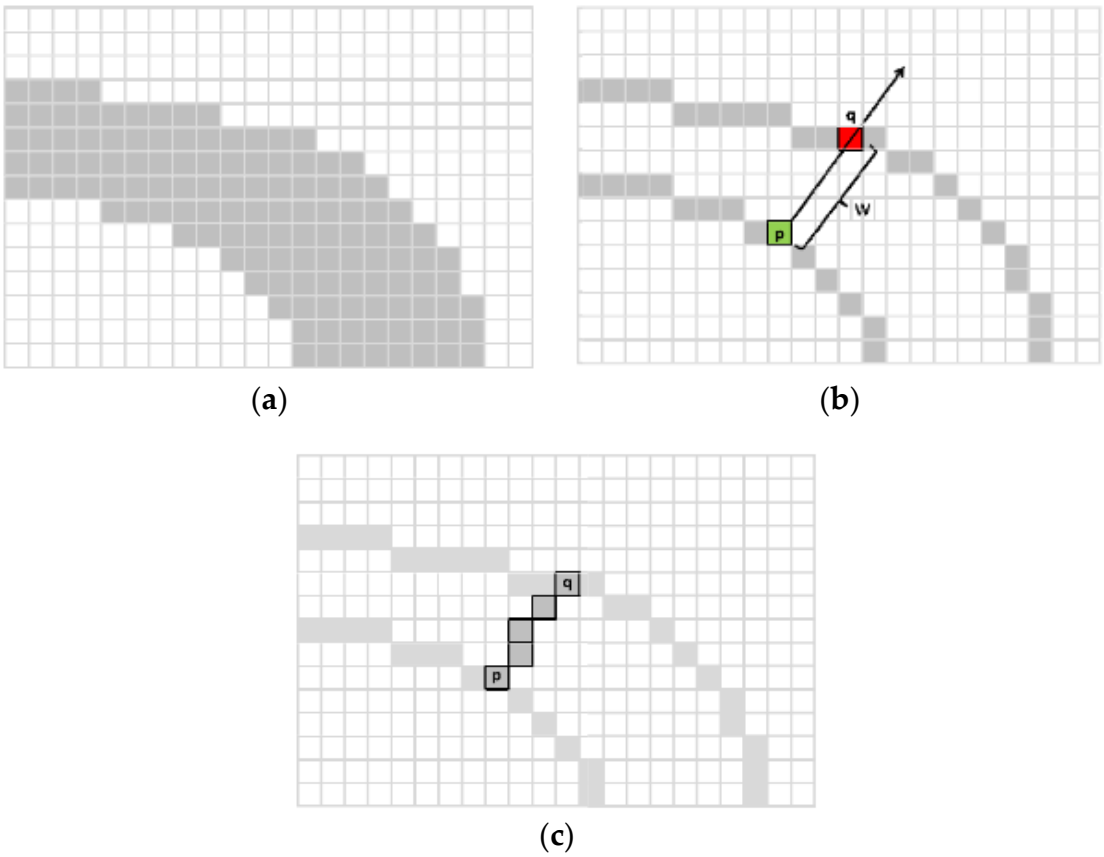

Figure 2. Implementation of the stroke width transform (SWT) [48]. (**a**) A typical stroke. (**b**) p is a pixel on the boundary of the stroke. (**c**) Each pixel along the ray is assigned by the minimum of its current value and the found width of the stroke.

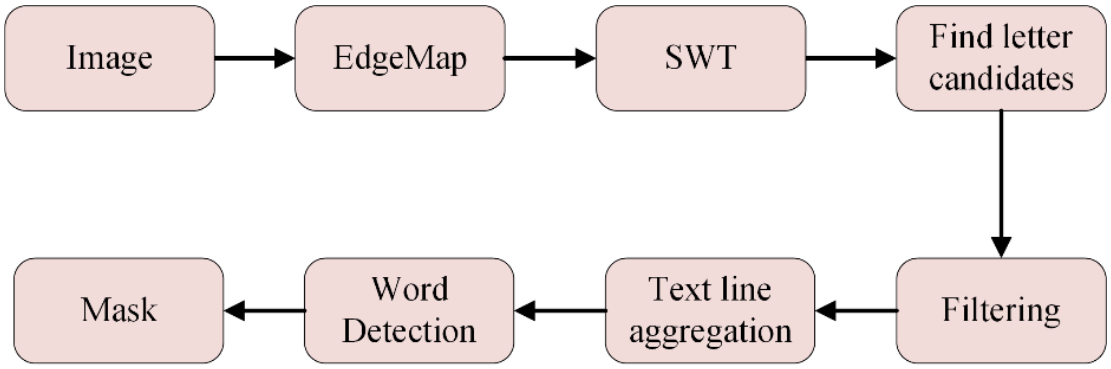

Figure 3. Detecting text in natural scenes text detection using the SWT [48].

In [49], the maximally stable extremal regions (MSER) [56] were leveraged to detect candidate text regions in scene images. This approach offers robustness to geometric, noise, and illumination conditions. The MSER method has been employed in several studies [16,57,58] that achieved excellent text detection performance on complex scene images. In [50], a text detection method based on color-enhanced contrast extremal regions (CERs) and neural networks was presented. The method used CERs to extract candidate text regions using a six-component tree which was produced from color images, and classify them into text regions and non-text regions using a neural network. In [59], a text detection method that combines extremal regions (ER) and corner-HOG features was presented. In [52], a novel method was proposed based on a conditional random field (CRF) pipeline, which used a convolutional neural network (CNN) to estimate the confidence of the MSER being a candidate text region. In [60,61], methods combining MSER and SWT were proposed to achieve better text detection performance.

It is worth mentioning that the scene text detection methods based on CCA mainly deal with extracting the connected regions in the image text candidate regions, thus reducing the search scope of natural scene text effectively. However, this type of method relies heavily on the detection of text-connected regions. In fact, in scene images with complex backgrounds, noise interference, low contrast, and color variation, it is difficult to detect connecting regions of text accurately. At the same time, it is also very difficult to design a reasonable analyzer for connected regions. In summary, CCA-based detection methods are difficult to implement and not robust in detecting text from scene images.

4. Deep Learning Approaches for STD

Deep learning methods automatically extract text features by training a model and are particularly suitable for object detection, speech recognition, and other pattern recognition problems. Typical deep learning networks include deep belief networks (DBN), convolutional neural networks (CNNs), recurrent neural networks (RNNs), and capsule networks. Deep learning-based methods are more effective, simple, and robust compared to manually designed algorithms for extracting and classifying candidate text regions. The boom of deep learning has also led to the development of successful techniques for scene text detection. In general, the main deep learning-based text detection methods can be classified into three categories: region proposal-based methods, image segmentation-based methods, and hybrid methods. Table 1 summarizes a comparison among some of the current state-of-the-art techniques in this field.

Table 1. Deep learning text detection methods, where S: Supervised, SS: Semi-Supervised, US: Un-Supervised, ST: Synthetic Text, IC13: ICDAR 2013, IC15: ICDAR 2015, M500: MSRA-TD500, IC17: ICDAR 2017 MLT, ToT: Total-Text, CTW: CTW-1500, COCO: COCO-Text.

Model_Based	Work	Source	Code	Backbone	Supervised			Training Datasets	Contributions
					S	SS	US	First-Stage	Fine-Tune
Convolutional Neural Network (CNN)	RSTD [16]	ECCV'14	-	-	✓	-	-	-	IC11, IC15
	DSOL [62]	ICLR'15	-	-	✓	-	-	MJSynth	-
	FOTS [63]	CVPR'18	✓	ResNet-50	✓	-	-	ST	IC13, IC15, IC17
	ABCNet [64]	CVPR'2020	✓	ResNet-50	✓	-	-	ST, COCO	ToT, CTW
	Corner [65]	CVPR'18	✓	VGG-16	✓	-	-	ST	IC13, 1C15
	DB [66]	AAAI'20	✓	ResNet	✓	-	-	ST	M500, CTW, ToT, IC15, IC17
Fully Convolutional Neural Network (FCN)	MOTD [13]	CVPR'16	✓	VGG-16	✓	-	-	-	IC13, IC15, M500
	EAST [67]	CVPR'17	✓	VGG-16	✓	-	-	-	IC15, COCO, M500
	STDH [68]	2016	✓	VGG-16	✓	-	-	-	IC13, IC15, M500
	E2ET [69]	CVPR'18	✓	PVANet [70]	✓	-	-	ST	IC13, IC15
	PixelLink [71]	AAAI'18	✓	VGG-16	✓	-	-	IC15	IC13, IC15, M500
	Textfield [72]	2019	✓	VGG-16	✓	-	-	ST	IC15, M500, ToT, CTW
Feature Pyramid Network (FPN)	PSENet [73]	AAAI'20	✓	ResNet	✓	-	-	IC17	IC13 or IC15
Faster R-CNN	CTPN [74]	ECCV'16	✓	VGG-16	✓	-	-	-	IC13
	R2CNN [75]	arXiv'17	✓	VGG-16	✓	-	-	IC15	-
	RRPN [76]	2018	✓	VGG-16	✓	-	-	M500	IC13, IC15
Mask-RCNN	MTSpotter [77]	ECCV'18	✓	ResNet-50	✓	-	-	ST	IC13, IC15, ToT
	PMTD [14]	CoRR'19	✓	ResNet-50	✓	-	-	IC17	IC13 or IC15
	MB [78]	2019	✓	ResNet-101	✓	-	-	ST	IC15, IC17, M500
Single Shot Detector (SSD)	SegLink [79]	CVPR'17	✓	VGG-16	✓	-	-	ST	IC13, IC15 or M500
	SSTD [80]	CVPR'17	✓	VGG-16	✓	-	-	-	IC13 or IC15
	TextBoxes++ [81]	2018	✓	VGG-16	✓	-	-	ST	IC15
	RRD [82]	CVPR'18	-	VGG-16	✓	-	-	ST	IC13, IC15, COCO, M500
U-Net	TextSnake [83]	ECCV'18	✓	VGG-16	✓	-	-	ST	IC15, M500, ToT, CTW
	CRAFT [84]	CVPR'19	✓	VGG-16	-	✓	-	ST	IC13, IC15, IC17

4.1. Region Proposal-Based Methods

Region proposal-based text detection methods adopt a general object detection framework, often using regression text boxes to obtain regional text information. In [85], a method based on a region proposal mechanism for text detection was proposed, while in [86], the Faster R-CNN [87] was improved using the inception structure proposed by GoogleNet. This resulted in the inception region proposal network (InceptionRPN), which obtains text candidate regions, uses a text detection network to remove background regions, and finally, votes on the overlapping detected regions to obtain the best results. In [75], a new method called the rotational region CNN (R2CNN) was proposed for detecting arbitrarily-oriented text in scene images. In [74], a novel connectionist text proposal network (CTPN) was proposed for localizing text lines in scene images, while a vertically-regressed proposal network (VRPN) was proposed in [88], which could match text regions using multiple neighboring small anchors. In [76], the rotation region proposal network (RRPN) was proposed to detect arbitrarily-oriented text by rotating text region proposals.

Rectangle bounding boxes or quadrangles have been adopted to describe text. In [89], an end-to-end trainable one-stage algorithm similar to a single shot multibox detector (SSD) [90] was proposed. Reference [81] was also based on an SSD object detection framework, where the rectangular box representation of conventional object detectors was replaced using a quadrilateral or rotated rectangle representation. In [91], a quadrilateral window (not a rectangle) was used to detect text in arbitrary orientations. A quadrilateral region proposal network (QRPN) was proposed in [92] for generating quadrilateral proposals based on a novel quadrilateral regression algorithm. In [82], two separate network branches were used to extract different text characteristics for text detection and oriented bounding box regression. In [93], corners were employed to estimate the possible locations of text instances, while an embedded data augmentation module inside a region-wise subnetwork was employed.

To achieve high coverage of the target box, in [94], a learning mechanism was proposed that integrates a two-stage R-CNN framework into a single-stage detector and uses the learned anchors instead of the original anchors into the final prediction. Existing deep neural network-based text detection methods use multi-scale filters and feature layers to detect multi-scale text. Reference [95] proposed a text detector named the short path network (SPN) to use low-level semantic features to complement the propagated loss of deep features. In [96], a multi-scale shape regression network (MSR) was presented, which was capable of locating text lines of different lengths, shapes, and curvatures in scenes. A scale-insensitive adaptive region proposal network (Adaptive-RPN) was proposed in [97] to generate text proposals by focusing only on the intersection over union (IoU) values between predicted and ground-truth bounding boxes. A scale-transfer module and a scale-relationship module were proposed in [98] to handle the problem of scale variation. A novel text detector, namely Look More than Once (LOMO), was presented in [99] to detect long text and arbitrarily shaped text in scene images by considering the geometric properties of the text instance, including the area, text center line, and border offsets to identify irregular text.

When trained with rigid word-level bounding boxes, the abovementioned methods exhibit limitations in analyzing text regions of an arbitrary shape. In [100], a one-stage model named convolutional character network (CharNet) was proposed, which predicts the bounding boxes of words and characters directly. In [101], a two-stage method called omnidirectional pyramid mask proposal text detector (OPMP) was proposed, which uses an effective pyramid sequence modeling method to produce arbitrary-shaped proposals. In [64], a novel adaptive Bezier-curve network (ABCNet) was presented to detect arbitrarily-shaped text in scene images.

This group of methods usually includes two parts: classification and regression of text candidate regions. In one-stage detectors, these candidate regions are generated by sliding windows; in two-stage detectors, the candidate regions are proposals generated by an RPN, but the RPN itself still classifies and regresses proposals generated by sliding windows. To improve the accuracy of text detection, it is often necessary to manually design anchors of various scales, aspect ratios, and even orientations to

better surround the text area, which makes region proposals-based methods complicated and inefficient. The anchor mechanism is not effective enough for scene text detection, which can be attributed to its IoU-based matching criterion between anchors and ground-truth boxes.

4.2. Image Segmentation-Based Methods

Image segmentation is an important part of image processing and machine vision techniques for image analysis and is a hot research topic today. Image segmentation is to classify images at the pixel level, determine the category of each point, and divide the image area. It is currently widely used in medical imaging, automated driving, UAV assistance, remote sensing, and other applications. Scene text detection can also be regarded as a pixel-level text/non-text classification, so image segmentation algorithms, such as semantic segmentation and instance segmentation, can be used to handle this challenge. In [13,68], an image segmentation-based method was proposed that used a fully convolutional network (FCN) [102] for pixel-level multi-oriented text detection. In [93], an algorithm for word-level text detection consisting of two cascaded CNNs was presented. The first network was fully convolutional and responsible for detecting regions containing text, while the second network predicted directional rectangles containing single word regions. A novel progressive scale expansion network (PSENet) approach was presented in [103], which gradually expanded the detection region from small kernels to large, and complete text instances were detected through multiple semantic segmentation maps. The system is robust, being able to detect arbitrarily shaped text and independently attached text. In [79], the segment linking (SegLink) method was introduced, which decomposed text into two components, namely segments and links. Instance aware component grouping (ICG) for arbitrary-shape text detection was presented in [104], while in [71], an instance segmentation-based method was proposed, which predicted text instances lying very close by linking pixels within the same instance. In [84], a new scene text detection method was proposed to detect text area effectively by exploring each character and affinity between characters. In [105], a mask R-CNN-based text detector was used to suppress false detection caused by background noise more effectively using the pyramid attention network (PAN) as a new backbone network. A novel framework with the local segmentation network (LSN) was presented in [106], followed by the curve connection to detect text in horizontal, oriented, and curved forms. An efficient and accurate arbitrary-shaped text detector, named the pixel aggregation network (PAN), was proposed in [107], which was equipped with a segmentation head made up of a feature pyramid enhancement module (FPEM) and a Feature Fusion Module (FFM). The authors of this study proposed a method based on mask R-CNN, named pyramid mask text detector (PMTD) [14], which used location-aware information to generate text masks instead of binary text masks. In [66], a module named differentiable binarization (DB) was proposed, which could perform the binarization process in a segmentation network. An FCN-based method named TextEdge was proposed in [108], where the text-region edge map was used as a segmentation mask. A segmentation-based detector named instance segmentation network (ISNet) was introduced [12], which linearly combines a generation mask and mask coefficients for fast text localization. A segmentation-based method that used polygon offsetting combined with border augmentation to detect text in natural images was presented in [109], while in [110], a novel character candidate extraction method based on super-pixel segmentation and hierarchical clustering was introduced. A novel scene text detection technique making use of semantics-aware text borders and bootstrapping-based text segment augmentation was presented in [111]. In [112], an instance segmentation-based framework was presented, which extracted each text instance as a separate connected component and introduced a shape-aware loss of adaptive multi-scale text instances when training the detection model.

Image segmentation-based methods for text/non-text classification at the pixel level have become mainstream for detecting text with multiple orientations and arbitrary shapes. However, this group of methods often requires time-consuming and complex post-processing to deal with complicated cases such as sticking or overlapping text.

4.3. Hybrid Methods

To detect scene text under more complex situations more efficiently, some researchers have combined the previous methods. In [113], a novel anchor-free region proposal network (AF-RPN) was proposed to replace the original anchor-based RPN and speed up text detection. In [114], a new framework for text detection named "Simple but Accurate" (SA-Text) was introduced, which utilizes heatmaps to detect text regions in scene images effectively. SA-Text detects text that occurs in various fonts, shapes, and orientations in scene images with complicated backgrounds. In [67], a new pipeline that directly predicts arbitrary orientations and quadrilateral text or text lines from natural images through a single network was proposed, eliminating unnecessary post-processing. A pixel-wise method named TextCohesion for scene text detection was proposed in [115], which splits a text instance into five key components: a text skeleton and four directional pixel regions. A novel conditional spatial expansion (CSE) mechanism to improve the performance of text detection by using a region expansion algorithm was introduced in [116]. CSE starts with a seed arbitrarily initialized within a text region and progressively merges neighborhood regions based on local features extracted by a CNN and contextual information of merged regions.

The advantages and disadvantages of the three kinds of methods for text detection are summarized in Table 2.

Table 2. Comparison of different kinds of text detection methods.

Method	Strength	Weakness
Region proposal-based	Higher detection accuracy and recall rates	Rely on complex frame designs and computationally intensive
Image segmentation-based	Be insensitive to font variation, noise, blur, and orientation	Weak detection of sticky or overlapping text
Hybrid methods	Can handle arbitrary strings and is robust to detection	Need to design innovative detection frameworks

5. STD Resources: Datasets

High-quality data and textual annotations are essential for both model learning and evaluation. Below, we summarize the most widely used benchmark datasets. A comprehensive list is provided in Tables 3 and 4.

Table 3. Benchmark datasets overview.

Datasets	Year	Language	URL
SVT	2010	English	http://vision.ucsd.edu/~kai/grocr/
KAIST	2011	Multi-lingual	http://www.iapr-tc11.org/mediawiki/index.php?title=KAIST_Scene_Text_Database
IC11	2011	English	http://www.cvc.uab.es/icdar2011competition/?com=downloads
M500	2012	English/Chinese	http://pages.ucsd.edu/~ztu/Download_front.htm
IC13	2013	English	http://dagdata.cvc.uab.es/icdar2013competition/?ch=2&com=downloads
USTB-SV1K	2015	English	http://prir.ustb.edu.cn/TexStar/MOMV-text-detection/
IC15	2015	English	http://rrc.cvc.uab.es/?ch=4&com=downloads
COCO-Text	2016	English	https://vision.cornell.edu/se3/coco-text-2/
SynthText	2016	English	http://www.robots.ox.ac.uk/~vgg/data/scenetext/
CTW	2017	Chinese	https://ctwdataset.github.io/
RCTW-17	2017	English/Chinese	http://rctw.vlrlab.net/dataset/
ToT	2017	English	https://github.com/cs-chan/Total-Text-Dataset
CTW	2017	English/Chinese	https://github.com/Yuliang-Liu/Curve-Text-Detector
MLT17	2017	Multi-lingual	http://rrc.cvc.uab.es/?ch=8
ArTs19	2019	English/Chinese	https://rrc.cvc.uab.es/?ch=14
MLT19	2019	Multi-lingual	https://rrc.cvc.uab.es/?ch=15&com=downloads
LSVT19	2019	English/Chinese	https://rrc.cvc.uab.es/?ch=16

Table 4. Details of benchmark datasets.

Datasets	Total	Train	Test	Arbitrary-Shape	Multi-Oriented	Annotation		
						Char	Word	Text-Line
IC03	509	258	251	×	×	×	✓	×
SVT	350	100	250	✓	×	✓	✓	×
KAIST	3000	-	-	×	×	×	✓	×
IC11	484	229	255	×	×	✓	✓	×
M500	500	300	200	✓	×	×	×	✓
IC13	462	229	233	×	×	✓	✓	×
USTB-SV1K	1000	500	500	✓	×	×	✓	×
IC15	1500	1000	500	✓	×	×	✓	×
COCO-Text	63,686	43,686	20,000	✓	×	×	✓	×
SynthText	85,8750	-	-	×	×	✓	✓	×
CTW	32,285	25,887	6398	✓	×	✓	✓	×
RCTW-17	12,514	15,114	1000	✓	×	×	×	✓
ToT	1525	1225	300	✓	✓	×	✓	✓
CTW	1500	1000	500	✓	✓	×	✓	✓
MLT17	18,000	7200	10,800	✓	×	×	✓	×
ArTs19	10,166	5603	4563	✓	✓	×	✓	×
MLT19	20,000	10,000	10,000	✓	×	×	✓	×
LSVT19	450,000	430,000	20,000	✓	✓	×	✓	✓

ICDAR 2003 (IC03) [117]. This is the first benchmark released for scene text detection and recognition from the ICDAR Robust Reading Competition. There are 258 natural images for training and 251 natural images for testing. All the text instances in this dataset are in English and are horizontally placed.

Street View Text (SVT) [27]. This dataset consists of 350 images annotated with word-level axis-aligned bounding boxes from Google Street View. It contains smaller and lower resolution text, and not all text instances within it are annotated.

KAIST [118]. This dataset comprises 3000 images captured in different environments, including outdoor and indoor scenes, under different lighting conditions (clear day, night, strong artificial lights, etc.). The images were captured either using a high-resolution digital camera or a low-resolution mobile phone camera. All images have been resized to 640×480 pixels.

ICDAR 2011 [119]. This dataset inherits from ICDAR 2003 and includes some modifications. There are 229 scene images for training and 255 scene images for testing.

MSRA-TD500 (M500) [120]. This dataset contains 500 natural scene images in total, with 300 images intended for training and 200 images for testing. It provides text-line-level annotation and polygon boxes for text region annotation. It contains both English and Chinese text instances.

ICDAR 2013 (IC13) [121]. This is also a modified version of ICDAR 2003. There are 229 natural images for training and 233 natural images for testing.

USTB-SV1k [122]. It contains 1000 street images from Google Street View with 2955 text instances in total. It provides word-level annotations, and it only considers English words.

ICDAR 2015 (IC15) [25]. It contains 1500 scene images, 1000 for training and 500 for validation/testing. The text instances (annotated using four quadrangle vertices) are usually skewed or blurred since they were acquired without users' prior preference or intention. Specifically, it contains 17,548 text instances. It provides word-level annotations. IC15 is the first incidental scene text dataset, and it is an English dataset.

COCO-Text [123]. It is the largest benchmark that can be used for text detection and recognition so far. The original images are from the Microsoft COCO dataset, and 173,589 text instances from 63,686 images are annotated. There are 43,686 images for training and 20,000 images for testing, including handwritten and printed, clear and blurry, English and non-English text.

SynthText [124]. It contains 858,750 synthetic images, where text with random colors, fonts, scales, and orientations are rendered on-scene images carefully to have a realistic look. The text in this dataset is annotated at the character, word, and line level.

Chinese Text in the Wild (CTW) [125]. This dataset contains 32,285 high-resolution street view images annotated at the character level, including its underlying character type, bounding box, and detailed attributes, such as whether word art has been used. The dataset is the largest one to date and the only one that contains detailed annotations. However, it only provides annotations for Chinese text and ignores other scripts, e.g., English. It contains 32,285 high-resolution street view images of Chinese text, with 1,018,402 character instances in total. All images are annotated at the character level, including its underlying character type, bounding box, and six other attributes. These attributes indicate whether the background is complex, whether it is raised, whether the text is hand-written or printed, occluded, and distorted, or whether word art has been used.

RCTW-17 [126]. This dataset contains various kinds of images, including street views, posters, menus, indoor scenes, and screenshots for competition on detecting and recognizing Chinese text in images. The dataset contains about 8000 training images and 4000 test images, with annotations similar to ICDAR2015.

Total-Text (ToT) [127]. This dataset contains a relatively large proportion of curved text, compared to the few instances in the previous datasets. These images were mainly obtained from street billboards and annotated as polygons with a variable number of vertices.

SCUT-CTW1500 [128]. This dataset contains 1500 images in total, 1000 for training and 500 for testing, with 10,751 cropped word images. Annotations in CTW-1500 are polygons with 14 vertices. The dataset mainly consists of Chinese and English words.

ICDAR 2017 MLT (MLT17) [129]. It is a large-scale multi-lingual text dataset, which contains 10,000 natural scene images in total, with 7200 training images, 1800 validation images, and 9000 testing images. It provides word-level annotation.

ICDAR 2019 Arbitrary-Shaped Text (ArT19) [130]. ArT consists of 10,166 images, 5603 for training, and 4563 for testing. They were collected with text shape diversity in mind, and all text shapes (i.e., horizontal, multi-oriented, and curved) have a high number of instances.

ICDAR 2019 MLT (MLT19) [131]. This dataset contains 18,000 images in total, with word-level annotation. Compared to MLT, this dataset has 10 languages. It is a more real and complex dataset for scene text detection.

ICDAR 2019 Large-scale Street View Text (LSVT19) [132]. This dataset consists of 20,000 testing images and 30,000 training images with full annotations, and 400,000 training images with weak annotations, which are referred to as partial labels.

6. Evaluations

6.1. Evaluation Metrics for STD

In this section, we summarize evaluation protocols for text detection algorithms. The task of text detection is commonly evaluated using the ICDAR protocol, the AP-based protocol, and the TIoU-metric, analyzed in the following paragraphs. The evaluation methods mainly consider three performance parameters, namely, Precision (P), Recall (C), and the overall evaluation index (F-measure, F). Commonly, recall and precision are calculated before F-mean, while there are some differences in the calculation methods. Determining whether two bounding boxes match or not is a straightforward but not simple problem. There are four ways in which two bounding boxes can match, as shown in Figure 4.

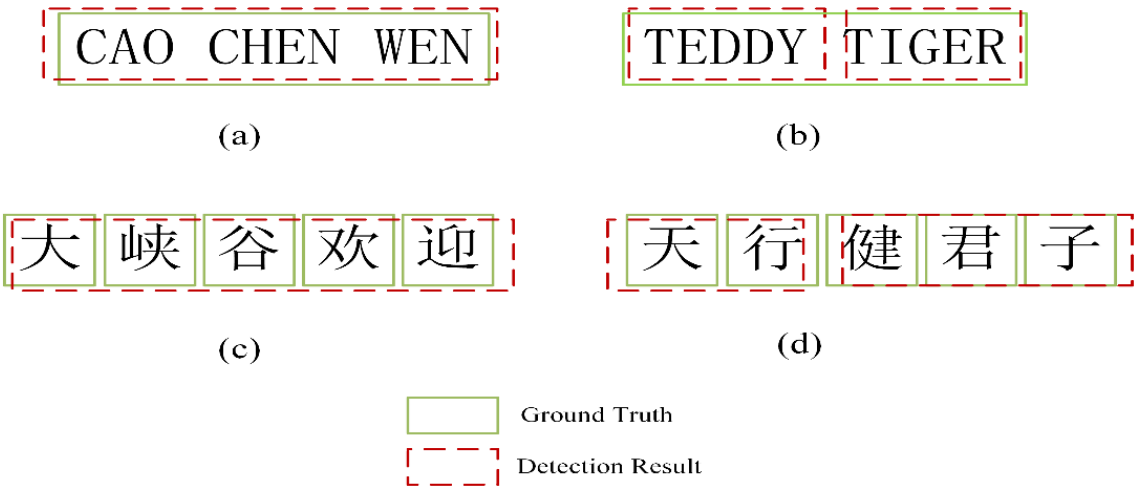

Figure 4. Four match types between ground truth and detected rectangles: (**a**) one-to-one match; (**b**) one-to-many matches with one ground truth rectangle; (**c**) one-to-many matches with one detection rectangle; (**d**) many to many matches.

6.1.1. ICDAR Evaluation Protocols

ICDAR 2003(IC03) Evaluation metrics. We have a set of ground-truth targets G, and the set of detection targets D. The IC03 metric calculates precision, recall, and the standard F-measure for one-to-one matches, as shown in Figure 4a, as follows:

$$Recall(G,D) = \frac{\sum_{i=1}^{|G|} BestMatch_G(G_i)}{|G|} \tag{1}$$

$$Precision(G,D) = \frac{\sum_{j=1}^{|D|} BestMatch_D\left(D_j\right)}{|D|} \tag{2}$$

We adopt the standard F-measure to combine precision and recall into a single measure of quality. The relative weights of these are controlled by α, which we set to 0.5 to give equal weight to precision and recall:

$$f = \frac{1}{\alpha/p' + (1-\alpha)/r'} \tag{3}$$

where $BestMatch_G$ and $BestMatch_D$ indicate the result of the closest match between the detection and ground truth rectangles, as defined below:

$$BestMatch_G(G_i) = \max_{j=1\ldots|D|} \frac{2\,Area\left(G_i \cap D_j\right)}{Area(G_i) + Area\left(D_j\right)} \tag{4}$$

$$BestMatch_D\left(D_j\right) = \max_{i=1\ldots|G|} \frac{2\,Area\left(D_j \cap G_i\right)}{Area\left(D_j\right) + Area(G_i)} \tag{5}$$

ICDAR 2013(IC13) Evaluation Metric. IC03 only considers one-to-one match types, which is simple but cannot handle all the cases detected, so in IC13, a new evaluation method was used: DetEval. The new method takes into account one-to-one, one-to-many, and many-to-one cases but does

not handle many-to-many cases. The criteria of these two evaluations are based on mutual overlap rates between detection ({D_j}j) and ground truth ({G_i}i):

$$\frac{A\left(G_i \cap D_j\right)}{A\left(D_j\right)} > tp \tag{6}$$

$$\frac{A\left(G_i \cap D_j\right)}{A(G_i)} > tr \tag{7}$$

where tp and tr are the thresholds of precision and recall, respectively.

$$P' = \frac{\sum_i Match_D\left(D_i, G, t_r, t_p\right)}{|D|} \tag{8}$$

$$R' = \frac{\sum_j Match_G\left(G_j, D, t_r, t_p\right)}{|D|} \tag{9}$$

where $Match_D$ and $Match_G$ are functions that consider the different types of matches:

$$Match_D\left(D_i, G, t_r, t_p\right) = \begin{cases} 1 \text{ if } D_i \text{ matches against a single detected rectangle} \\ 0 \text{ if } D_i \text{ does not match against any detected rectangle} \\ f_{sc}(k) \text{ if } D_i \text{ matches against several } (\rightarrow k) \text{ detected rectangles} \end{cases} \tag{10}$$

$$Match_G\left(G_i, D, t_r, t_p\right) = \begin{cases} 1 \text{ if } G_j \text{ matches against a sin gle detected rectangle} \\ 0 \text{ if } G_j \text{ does not match against any detected rectangle} \\ f_{sc}(k) \text{ if } G_j \text{ matches against several } (\rightarrow k) \text{ detected rectangles} \end{cases} \tag{11}$$

where $f_{sc}(k)$ is a parameter function that controls the amount of punishment, and it is often set to 0.8.

ICDAR 2015(IC15) IoU Metric. The IC15 metric [8] follows the same metric as the Pascal VOC. To be considered a correct detection, the value of Intersection-over-Union (IoU) defined in equation 12 must exceed 0.5.

$$\frac{A\left(G_j \cap D_i\right)}{A\left(G_j \cup D_i\right)} > 0.5 \tag{12}$$

6.1.2. AP-Based Evaluation Methods

To avoid fine-tuning the output detection confidence, datasets, such as RCTW-17 [26], have adopted interpolated average precision as the main detection evaluation metric: For a given task and class, the precision-recall curve is computed based on the method's ranked output. Basically, this metric relies on the IoU metric to calculate the precision and recall in advance.

6.1.3. Tightness-Aware Intersection-Over-Union (TIoU) Evaluation Protocol

The existing metrics exhibit some drawbacks: (1) They are not goal-oriented; (2) they cannot recognize the tightness of detection methods; (3) existing one-to-many and many-to-one solutions involve inherent loopholes and deficiencies. Previous metrics severely rely on an IoU threshold, which will lead to unreasonable results in some cases, such as those shown in Figure 5. To improve these shortcomings, the TIoU approach involves three annotation concepts to enhance the focus on detecting text content: (a) The annotation does not cut the text instance; (b) annotation contains less background noise, especially outlier text instances; (c) even if annotations do not match the text instance perfectly, they should be as perfect as possible.

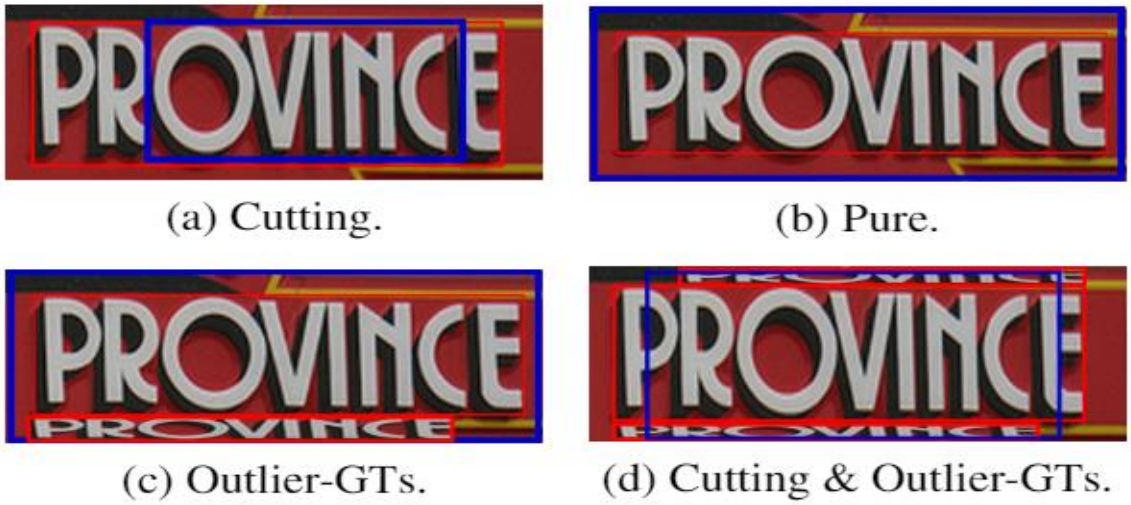

Figure 5. (**a**–**d**) all have the same detection (blue) intersection over union (IoU) of 0.66 against the GT (Ground Truth) (red).

The text instances detected in Figure 5a,b have the same value of IoU (0.66) against the ground truth, while the former does not detect a few characters of the GT (Ground Truth). To solve this issue, the cutting behavior can be penalized using the corresponding proportion of intersection in GT, as shown in Equations (13) and (14):

$$\mathrm{TIoU}_{\mathrm{Recall}} = \frac{A\left(G_i \cap D_j\right) * f(C_t)}{A\left(G_i \cup D_j\right)} \tag{13}$$

$$f(C_t) = 1 - x, x = \frac{C_t}{A(G_i)} \tag{14}$$

The proposed solution aims to penalize such types of detections for making detection compact for avoiding including outlier-GTs in the same detected region. Nevertheless, as shown in Figure 5c, if the outlier-GTs are inside the target GT region, even the perfect detection bounding box cannot avoid containing these outliers. Therefore, only the outlier-GT region that is inside the detection bounding box but outside the target GT region would be penalized. The area (O_t) of the union of all eligible outlier-GTs is calculated using Equation (15):

$$\mathrm{TIoU}_{\mathrm{Precision}} = \frac{A\left(D_j \cap G_i\right) * f(O_t)}{A\left(D_j \cup G_i\right)} \tag{15}$$

$$f(O_t) = 1 - x, x = \frac{O_t}{A\left(D_j\right)} \tag{16}$$

6.1.4. Discussion

In this part, we briefly summarize the strengths and drawbacks of commonly used evaluation methods for scene text detection. Details are shown in Table 5.

Table 5. Evaluation methods for scene text detection.

Evaluation Protocols	Match Type	Strength and Weakness
IC03 Evaluation Protocol	One-to-One	The IC03 metric calculates precision, recall, and the standard F-measure for one-to-one matches. However, it is unable to handle one-to-many and many-to-many matches between the ground truth and detections.
IC13 Evaluation Protocol	One-to-One One-to-Many Many-to-one	This method takes into account one-to-one, one-to-many, and many-to-one cases but cannot handle many-to-many cases.
IC15 Evaluation Protocol	One-to-One One-to-Many Many-to-one	This method uses the ICDAR15 intersection over union (IoU) metric and IoU ≥ 0.5 as a threshold for counting a correct detection. This method is the most commonly used evaluation method and is simple to calculate.
TIoU Evaluation Protocol	One-to-one One-to-many Many-to-one many-to-many	This method can quantify the completeness of ground truth, the compactness of detection, and the tightness of the matching degree. However, it is relatively complex to calculate.

6.2. Results on Benchmark Datasets

In this section, we present the results of representative text detection methods on some public datasets. The evaluation uses the Precision (P), Recall (C), and F-measure (F) metrics. Since different methods may involve experiments on different benchmark datasets, and even on the same dataset they may adopt different training sets (such as using a synthetic dataset for pre-training or using special data augmentation schemes to enlarge the number of training samples), it is impossible to make an absolutely fair comparison. However, the analysis is useful for evaluating the development of state-of-the-art methods in this field and establishing future directions.

Table 6 reports the text detection performance of different methods on a horizontal-text dataset. As is shown in Table 6, on the IC13 dataset, the performance has increased drastically from 77% ([46]) to 92.1% ([133]) in terms of F-measure. In [133], a supervised pyramid context network (SPCNET) (see Figure 6) is adopted, which can achieve better detection results. It can be observed that multiple methods of general object detection and semantic segmentation have been extended to scene text location, and the current trend is applying a deep learning framework to training an end-to-end text detector.

Table 6. Horizontal text detection results. Dataset: IC13.

Work	Source	P	R	F
FASText [46]	CVPR'15	84	69	77
FCRN [124]	CVPR'16	93.8	76.4	84.2
CTPN [74]	ECCV'16	93	83	88
WordSup [134]	CVPR'17	93.3	87.5	90.3
DMPNet [91]	CVPR'17	93	83	87
ArbiText [135]	2017	82.6	93.6	87.7
SSTD [80]	CVPR'17	89	86	88
SegLink [79]	CVPR'17	87.7	83	85.3
RTN [136]	IAPR'17	94	89	91
EAST [67]	CVPR'17	93	83	87
AF-RPN [113]	ICDAR'17	94	90	92
PixelLink [71]	AAAI'18	88.6	87.5	88.1
MCN [137]	2018	88	87	88
Border [111]	2018	91.5	87.1	89.2
TextBoxes++ [81]	2018	92	86	89
RRPN [76]	2018	95	89	91
RGC [138]	ICIP'18	89	77	83
SPCNet [133]	AAAI'19	93.8	90.5	92.1
MSR [96]	2019	91.8	88.5	90.1
Roy et al. [139]	2020	90.4	88	89.1

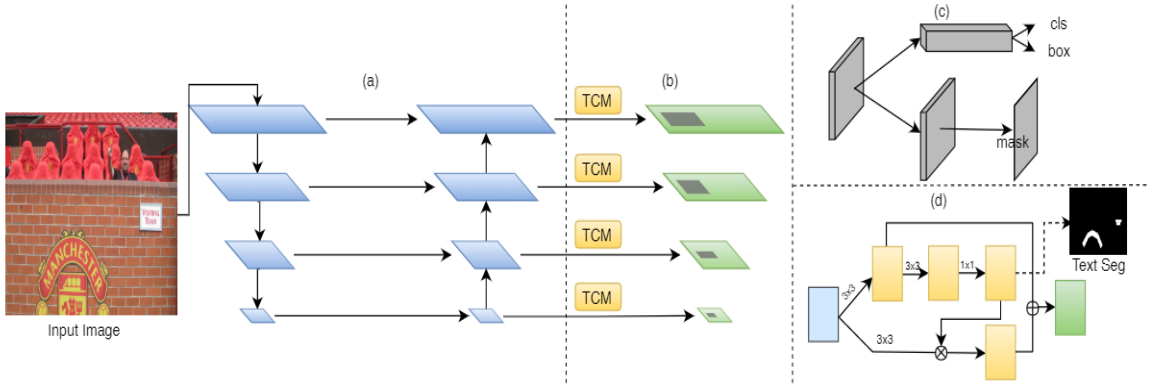

Figure 6. Supervised pyramid context network [133]. (**a**)The Feature Pyramid Network. (**b**) Pyramid Feature fusion. (**c**) Mask R_CNN branch. (**d**) The proposed Text-Context Module.

Table 7 shows the text detection performance of different methods on irregular-text datasets. The methods of [66,97] achieve relatively high performance, while in [97], ContourNet (Figure 7) was proposed to train an accurate arbitrarily-shaped text detection model. In [66], a segmentation-based network method was proposed, which can set the thresholds for binarization adaptively using a module named differentiable binarization (Figure 8).

Table 7. Irregular text detection results. Dataset: Total-Text (ToT) and Chinese Text in the Wild (CTW).

Work	Source	ToT			CTW		
		P	R	F	P	R	F
TextSnake [83]	ECCV'18	82.7	74.5	78.4	67.9	85.3	75.6
CRAFT [84]	CVPR'19	87.6	79.9	83.6	86	81.1	83.5
Liu et al. [116]	2019	81.4	79.1	80.2	78.7	76.1	77.4
Seglink++ [104]	2019	82.9	80.9	81.5	82.8	79.8	81.3
SAST [140]	2019	85.57	75.49	80.2	81.19	81.71	81.45
PAN [107]	CVPR'19	89.3	81	85	86.4	81.2	83.7
SPCNet [133]	AAAI'19	83	83	83	-	-	-
MSR [96]	2019	85.2	73	76.8	83.8	77.8	80.7
LOMO [99]	CVPR'19	87.6	79.3	83.3	-	-	-
SLPR [141]	2018	-	-	-	80.1	70.1	74.8
CTD+TLOC [128]	2019	-	-	-	77.4	69.8	73.4
PSENet [73]	AAAI'20	84	77.9	80.9	84.8	79.7	82.2
Tian et al. [112]	CVPR'19	-	-	-	81.7	84.2	80.1
Wang et al. [97]	2020	86.9	83.9	85.4	83.7	84.1	83.9
Roy et al. [139]	2020	88	79	83.25	85	82	83.47
OPMP [101]	2020	-	-	-	85.1	80.8	82.9
DB [66]	AAAI'20	87.1	82.5	84.7	86.9	80.2	83.4

Table 8 shows the text detection performance of different methods on arbitrary quadrilateral text datasets. As is shown in Table 8, [11] achieves relatively high performance on the IC15 dataset by applying the spatial binning positions in Position Sensitive ROI (PSROI) pooling (Figure 9). Besides, the method of ContourNet [97] achieves state-of-the-art performance on the M500 dataset.

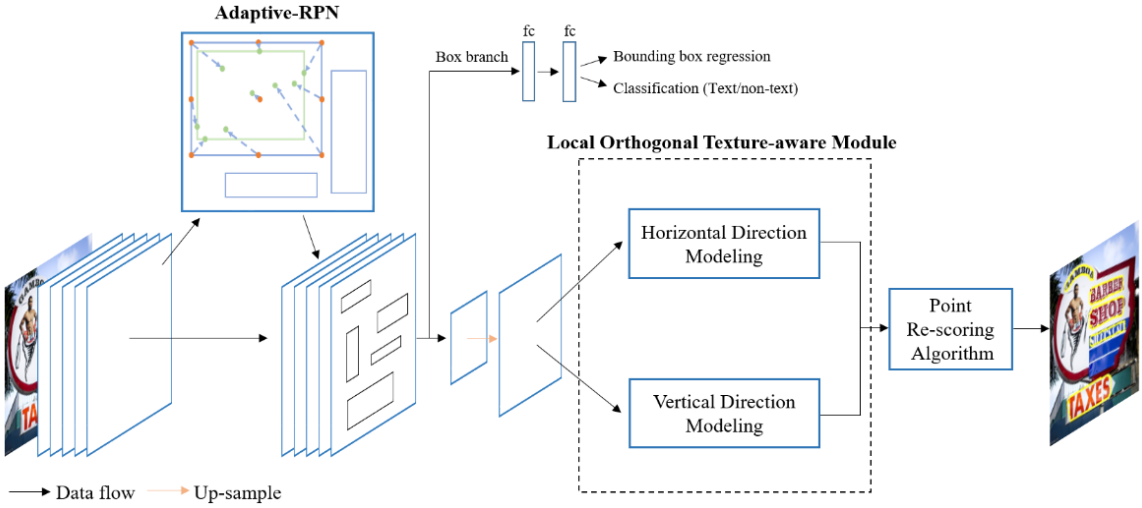

Figure 7. Pipeline of ContourNet [97].

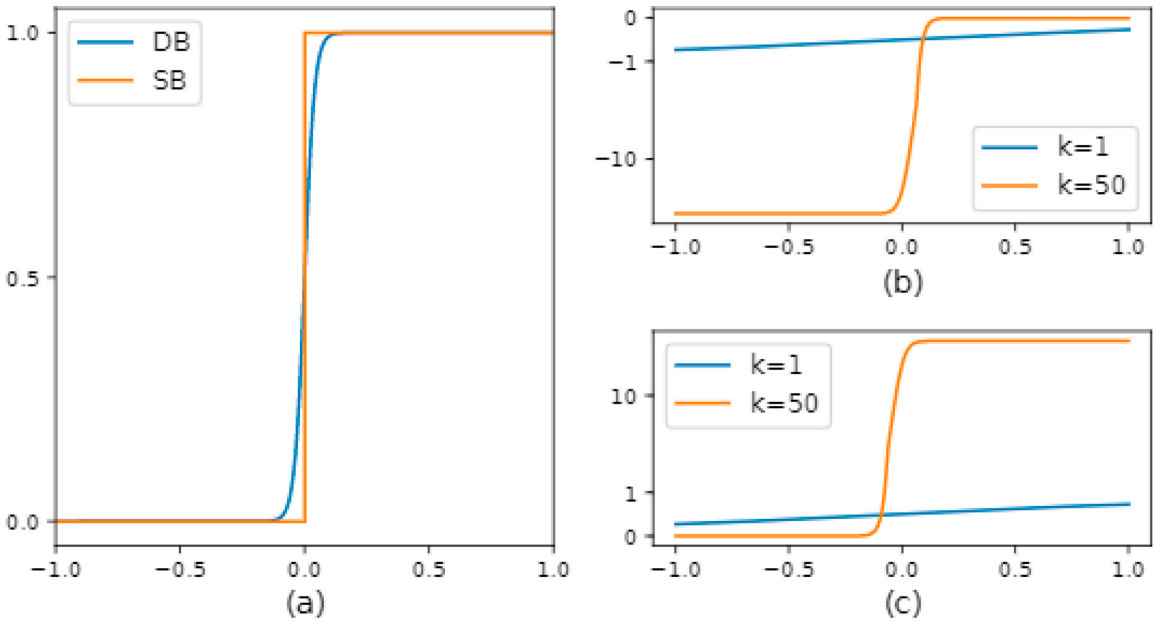

Figure 8. Differentiable binarization and its derivative. (**a**) Numerical comparison of standard binarization (SB) and differentiable binarization (DB). (**b**) Derivative of l+. (**c**) Derivative of l− [66].

Figure 9. Visualization of learned offset parts in deformable Position Sensitive ROI (PSROI) [11].

Table 8. Arbitrary quadrilateral text detection results. Dataset: M500 and IC15.

Work	Source	M500				IC15			
		P	R	F	FPS	P	R	F	FPS
Kang et al. [142]	2014	71	62	66	-	-	-	-	-
Zhang et al. [13]	2016	83	67	74	-	71	43	54	-
CTPN [74] [1]	ECCV'16	-	-	-	-	74.2	51.6	60.9	7.14
STDH [68]	2016	-	-	-	1.61	72.26	58.69	64.77	1.61
WordSup [134]	CVPR'17	-	-	-	-	79.3	77	78.2	2
SSTD [80]	2017	-	-	-	-	80	73	77	7.7
EAST-PVANet [67] [2]	CVPR'17	87.28	67.43	76.08	13.2	83.27	78.33	80.72	13.2
TextBoxes++ [81] [3]	2018	87.8	78.5	82.9	-	-	-	-	11.6
Border [111]	2018	78.2	58.8	67.1	-	-	-	-	-
RRD [96]	CVPR'18	59.1	77.5	67	10	-	-	-	6.5
TextSnake [83] [4]	ECCV'18	83.2	73.9	78.3	1.1	84.9	80.4	82.6	1.1
IncepText [11]	2018	87.5	79	83	3.7	93.8	87.3	90.5	3.7
Corner [65] [5]	CVPR'18	87.6	76.2	81.5	5.7	94.1	70.7	80.7	3.6
PixelLink [71] [6]	AAAI'18	-	-	-	3	85.5	82	83.7	3
PMTD [14] [7]	2019	-	-	-	-	91.3	87.4	89.3	-
CRAFT [84] [8]	CVPR'19	-	-	-	8.6	89.8	84.3	89.8	5.6
PSENet [73]	AAAI'20	-	-	-	-	86.9	84.5	85.7	-
LOMO [143]	2019	79.1	60.2	68.4	-	89	86	88	-
PuzzleNet [144]	2020	-	-	-	-	88.9	88.1	88.5	-
DB-ResNet18 [66] [9]	AAAI'20	90.4	76.3	82.8	62	84.8	77.5	81	55
DB-ResNet50 [66] [9]	AAAI'20	91.5	79.2	84.9	32	86.9	80.2	83.5	22
Roy et al. [139]	2020	88	78	82.6	-	91.1	83.1	86.9	-
OPMP [101]	2020	86	83.4	84.7	1.6	-	-	-	1.4
Dasgupta et al. [145]	2020	81.6	88.2	84.7	-	91.3	89.2	90.2	-
ContourNet [97]	CVPR'20	87.6	86.1	86.9	-	-	-	-	-

[1] http://textdet.com/, [2] https://github.com/argman/EAST, [3] https://github.com/MhLiao/TextBoxes_plusplus, [4] https://github.com/princewang1994/TextSnake.pytorch, [5] https://github.com/lvpengyuan/corner, [6] https://github.com/ZJULearning/pixel_link, [7] https://github.com/jjprincess/PMTD, [8] https://github.com/clovaai/CRAFT-pytorch, [9] https://github.com/MhLiao/DB.

Speedup: Current text detection methods place more emphasis on speed and efficiency, which is necessary for real-time scene text detection. As shown in Figures 10 and 11, DB [66] achieves state-of-the-art speed on MSRA-TD500 and ICDAR 2015 datasets. Specifically, with a backbone of ResNet-18, our detector achieves an F-measure of 82.8, running at 62 fps, on the MSRA-TD500 dataset, and running at 55 fps with an F-measure of 81 on the ICADAR 2015 dataset.

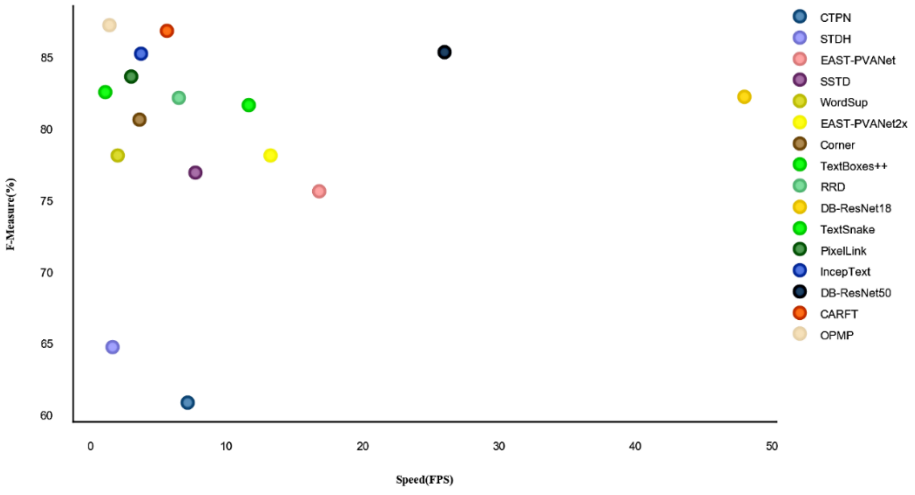

Figure 10. Detection speed of different methods on ICDAR-2015 dataset.

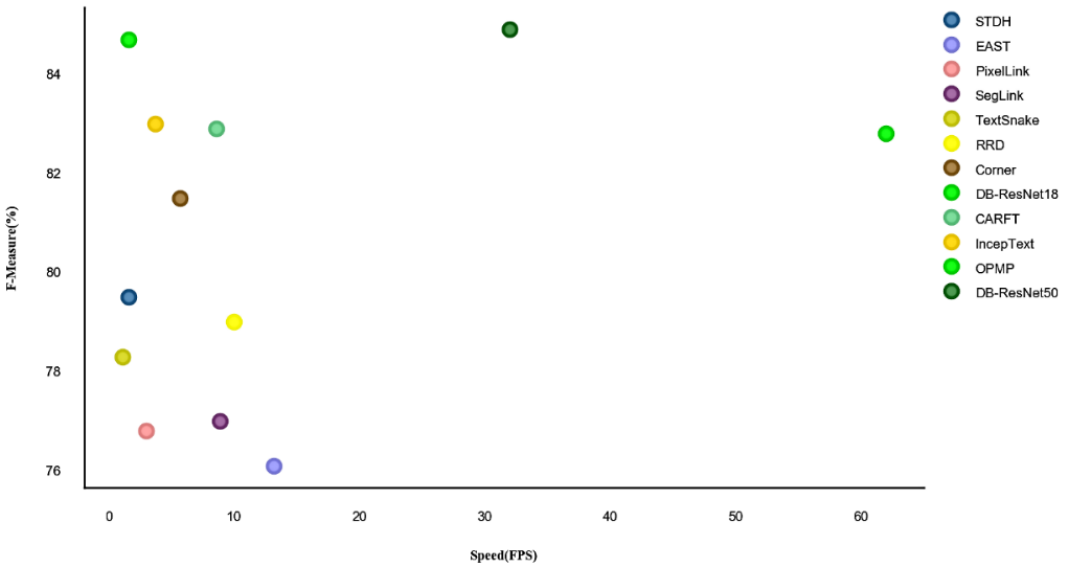

Figure 11. Detection speed of different methods on MSRA-TD500 dataset.

7. Conclusions and Discussion

In this paper, we reviewed scene text detection methods proposed in recent years. We comprehensively classified these methods into three types and highlighted the key techniques. Furthermore, we analyzed three types of benchmarks and evaluation protocols. Finally, we reported the results of several representative methods on benchmarks and compared their performance.

As discussed in Sections 3 and 4, from the manual design of text features to feature extraction using deep learning, DL-based STD significantly improved the speed as well as the accuracy of text detection. Deep learning-based methods for scene text detection have emerged with promising results. However, there are still some in the field.

Benchmark Dataset. Scene text detection frameworks, including deep learning-based STD methods, required large, annotated datasets for training. However, data annotation remains time-consuming and expensive. It is a big challenge to create a very large benchmark dataset, such as ImageNet, including plentiful scenarios, such as multi-scale, multi-lingual, and multi-orientation text, etc.

Real-time Scene Text Detection. Text information in scene images is extremely helpful to people's daily activities. Therefore, applying scene text detection technology to smart terminal devices (e.g., mobile phones, cameras, assistive devices, etc.) is a future development direction. However, most of the current methods are limited by the performance of smart terminal devices, which cannot achieve real-time levels while maintaining relatively good detection accuracy. Hence, real-time text detection is another future development direction.

Special Scene Text Detection Methods. Most of the proposed text detection methods mainly demonstrated their performance on some public datasets, and some of them simply accumulated some domain knowledge and adjusted parameters repeatedly (e.g., using Faster R-CNN, SSD, FCN, RNN, and other pattern recognition domain knowledge) to obtain higher testing performance, which leads to a lack of innovation and deep thinking. This results in approaches by researchers with no specialization in the field of document analysis.

Author Contributions: Original draft preparation, D.C., L.W.; Investigation, D.C., L.W., Y.H., J.D.; writing—review & editing, D.C., Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Intelligent analysis and control platform for high-risk operations, grant number 2020ZHZY0002.

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

References

1. Greenhalgh, J.; Mirmehdi, M. Recognizing Text-Based Traffic Signs. *IEEE Trans. Intell. Transp.* **2015**, *16*, 1360–1369. [CrossRef]
2. Yin, X.-C.; Zuo, Z.-Y.; Tian, S.; Liu, C.-L. Text Detection, Tracking and Recognition in Video: A Comprehensive Survey. *IEEE Trans. Image Process.* **2016**, *25*, 2752–2773. [CrossRef] [PubMed]
3. Ham, Y.K.; Kang, M.S.; Chung, H.K.; Park, R.-H.; Park, G.T. Recognition of raised characters for automatic classification of rubber tires. *Opt. Eng.* **1995**, *34*, 102. [CrossRef]
4. Shilkrot, R.; Huber, J.; Liu, C.; Maes, P.; Nanayakkara, S.C. FingerReader: A wearable device to support text reading on the go. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2014; pp. 2359–2364.
5. Hedgpeth, T.; Black, J.A., Jr.; Panchanathan, S. A demonstration of the iCARE portable reader. In Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility, New York, NY, USA, 23 October 2006; Volume 279.
6. Smith, R. A simple and efficient skew detection algorithm via text row accumulation. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; Volume 2, pp. 1145–1148. [CrossRef]
7. LEE, C.-M.; Kankanhalli, A. Automatice extraction of characters in complex scene images. *Int. J. Pattern Recogn.* **1995**, *9*, 67–82. [CrossRef]
8. Zhong, Y.; Karu, K.; Jain, A.K. Locating text in complex color images. *Pattern Recognit.* **1995**, *1*, 146–149. [CrossRef]
9. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
10. Richardson, E.; Azar, Y.; Avioz, O.; Geron, N.; Ronen, T.; Avraham, Z.; Shapiro, S. It's All about the Scale—Efficient Text Detection Using Adaptive Scaling. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision(WACV), Snowmass Village, CO, USA, 1–5 March 2020; pp. 1833–1842. [CrossRef]
11. Yang, Q.; Cheng, M.; Zhou, W.; Chen, Y.; Qiu, M.; Lin, W.; Chu, W. IncepText: A New Inception-Text Module with Deformable PSROI Pooling for Multi-Oriented Scene Text Detection. *arXiv* **2018**, arXiv:1805.01167.
12. Yang, P.; Yang, G.; Gong, X.; Wu, P.; Han, X.; Wu, J.; Chen, C. Instance Segmentation Network With Self-Distillation for Scene Text Detection. *IEEE Access* **2020**, *8*, 45825–45836. [CrossRef]
13. Zhang, Z.; Zhang, C.; Shen, W.; Yao, C.; Liu, W.; Bai, X. Multi-oriented Text Detection with Fully Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4159–4167.
14. Liu, J.; Liu, X.; Sheng, J.; Liang, D.; Li, X.; Liu, Q. Pyramid Mask Text Detector. *arXiv* **2019**, arXiv:1903.11800.
15. Christen, M.; Saravanan, A. RFBTD: RFB Text Detector. *arXiv* **2019**, arXiv:1907.02228.
16. Huang, W.; Qiao, Y.; Tang, X. Robust Scene Text Detection with Convolution Neural Network Induced MSER Trees. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 497–511.
17. Tang, Y.; Wu, X. Scene Text Detection and Segmentation Based on Cascaded Convolution Neural Networks. *IEEE Trans. Image Process.* **2017**, *26*, 1509–1520. [CrossRef]
18. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2015; Volume 2.
19. Shi, B.; Yang, M.; Wang, X.; Lyu, P.; Yao, C.; Bai, X. ASTER: An Attentional Scene Text Recognizer with Flexible Rectification. *IEEE Trans. Pattern. Anal.* **2018**, *41*, 2035–2048. [CrossRef] [PubMed]
20. Shi, B.; Wang, X.; Lyu, P.; Yao, C.; Bai, X. Robust Scene Text Recognition with Automatic Rectification. *arXiv* **2016**, arXiv:1603.03915.
21. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. Acm.* **2017**, *60*, 84–90. [CrossRef]
22. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv* **2016**, arXiv:1602.07261.

23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
24. Tychsen-Smith, L.; Petersson, L. DeNet: Scalable Real-time Object Detection with Directed Sparse Sampling. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 428–436.
25. Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; 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), Washington, DC, USA, 23–26 August 2015; pp. 1156–1160.
26. Gopalan, C.; Manjula, D. Sliding window approach based Text Binarisation from Complex Textual images. *arXiv* **2010**, arXiv:1003.3654.
27. Hutchison, D.; Kanade, T.; Kittler, J.; Kleinberg, J.M.; Mattern, F.; Mitchell, J.C.; Naor, M.; Nierstrasz, O.; Rangan, C.P.; Steffen, B.; et al. Word Spotting in the Wild. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 591–604.
28. Fabrizio, J.; Marcotegui, B.; Cord, M. Text detection in street level images. *Pattern Anal. Appl.* **2013**, *16*, 519–533. [CrossRef]
29. He, T.; Huang, W.; Qiao, Y.; Yao, J. Text-Attentional Convolutional Neural Network for Scene Text Detection. *IEEE Trans. Image Process.* **2016**, *25*, 2529–2541. [CrossRef]
30. Chen, X.; Yuille, A.L. Detecting and reading text in natural scenes. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 27 June–2 July 2004; Volume 2, pp. 366–373.
31. Lee, J.-J.; Lee, P.-H.; Lee, S.-W.; Yuille, A.; Koch, C. AdaBoost for Text Detection in Natural Scene. In Proceedings of the 2011 International Conference on Document Analysis and Recognition, Beijing, China, 18–21 September 2011; pp. 429–434.
32. Wei, Y.C.; Lin, C.H. A robust video text detection approach using SVM. *Expert Syst. Appl.* **2012**, *39*, 10832–10840. [CrossRef]
33. Zhang, Y.; Wang, C.; Xiao, B.; Shi, C. A New Method for Text Verification Based on Random Forests. In Proceedings of the 2012 International Conference on Frontiers in Handwriting Recognition, Bari, Italy, 18–20 September 2012; pp. 109–113.
34. Shi, C.; Wang, C.; Xiao, B.; Gao, S.; Hu, J. End-to-end scene text recognition using tree-structured models. *Pattern Recogn.* **2014**, *47*, 2853–2866. [CrossRef]
35. Li, H.; Doermann, D.; Kia, O. Automatic text detection and tracking in digital video. *IEEE Trans. Image Process.* **2020**, *9*, 147–156. [CrossRef]
36. Hanif, S.M.; Prevost, L. Text Detection and Localization in Complex Scene Images using Constrained AdaBoost Algorithm. In Proceedings of the 2009 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, 26–29 July 2009; pp. 1–5. [CrossRef]
37. Zhao, X.; Lin, K.-H.; Fu, Y.; Hu, Y.; Liu, Y.; Huang, T.S. Text from Corners: A Novel Approach to Detect Text and Caption in Videos. *IEEE Trans. Image Process.* **2011**, *20*, 790–799. [CrossRef] [PubMed]
38. Wu, W.; Chen, X.; Yang, J. Detection of Text on Road Signs from Video. *IEEE Trans. Intell. Transp. Syst.* **2005**, *6*, 378–390. [CrossRef]
39. Ye, Q.; Huang, Q.; Gao, W.; Zhao, D. Fast and robust text detection in images and video frames. *Image Vis. Comput.* **2005**, *23*, 565–576. [CrossRef]
40. Neumann, L.; Matas, J. Real-Time Scene Text Localization and Recognition. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, Rhode Island, 16–21 June 2012; pp. 3538–3545.
41. Mancas-Thillou, C.; Gosselin, B. Color text extraction with selective metric-based clustering. *Comput. Vis. Image Und.* **2007**, *107*, 97–107. [CrossRef]
42. Wang, K.; Kangas, J.A. Character location in scene images from digital camera. *Pattern Recogn.* **2003**, *36*, 2287–2299. [CrossRef]
43. Zhu, Y.; Yao, C.; Bai, X. Scene text detection and recognition: Recent advances and future trends. *Front. Comput. Sci.* **2015**, *10*, 19–36. [CrossRef]

44. Jamil, A.; Siddiqi, I.; Arif, F.; Raza, A. Edge-Based Features for Localization of Artificial Urdu Text in Video Images. In Proceedings of the 2011 International Conference on Document Analysis and Recognition, Beijing, China, 18–21 September 2011; pp. 1120–1124.
45. Liu, X.; Samarabandu, J. Multiscale Edge-Based Text Extraction from Complex Images. *IEEE Int. Conf. Multimed. Expo* **2006**, *52*, 1721–1724.
46. Buta, M.; Neumann, L.; Matas, J. FASText: Efficient Unconstrained Scene Text Detector. In Proceedings of the IEEE International Conference on Computer Vision, Las Condes, CL, USA, 11–18 December 2015; pp. 1206–1214.
47. Liu, C.; Wang, C.; Dai, R. Text detection in images based on unsupervised classification of edge-based features. In Proceedings of the Eighth International Conference on Document Analysis and Recognition (ICDAR'05), Seoul, Korea, 31 August–1 September 2005; Volume 2, pp. 610–614.
48. Epshtein, B.; Ofek, E.; Wexler, Y. Detecting text in natural scenes with stroke width transform. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2963–2970.
49. Neumann, L.; Matas, J. A Method for Text Localization and Recognition in Real-World Images. In Proceedings of the Asian Conference on Computer Vision, Queenstown, New Zealand, 8–12 November 2011; Springer: Berlin/Heidelberg, Germany, 2011; Volume 770–783.
50. Mosleh, A.; Bouguila, N.; Hamza, A.B. Image Text Detection Using a Bandlet-Based Edge Detector and Stroke Width Transform. In Proceedings of the BMVC, Guildford, UK, 3–7 September 2012; pp. 1–12. [CrossRef]
51. Huang, W.; Lin, Z.; Yang, J.; Wang, J. Text Localization in Natural Images Using Stroke Feature Transform and Text Covariance Descriptors. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1241–1248.
52. Karthikeyan, S.; Jagadeesh, V.; Manjunath, B.S. Learning bottom-up text attention maps for text detection using stroke width transform. In Proceedings of the 2013 IEEE International Conference on Image Processing, Melbourne, Australia, 15–18 September 2013; pp. 3312–3316.
53. Jameson, J.; Abdullah, S.N.H.S. Extraction of arbitrary text in natural scene image based on stroke width transform. In Proceedings of the 2014 14th International Conference on Intelligent Systems Design and Applications, Okinawa, Japan, 28–30 November 2014; pp. 124–128. [CrossRef]
54. Jian, H.; Xiaopei, L.; Qian, Z. A SWT Verified Method of Natural Scene Text Location. In Proceedings of the 2016 International Symposium on Computer, Consumer and Control (IS3C), Xi'an, China, 4–6 July 2016; pp. 980–984. [CrossRef]
55. Titijaroonroj, T. Modified Stroke Width Transform for Thai Text Detection. In Proceedings of the 2018 International Conference on Information Technology (InCIT), Khon Kaen, Thailand, 24–25 October 2018; pp. 1–5.
56. Matas, J.; Chum, O.; Urban, M.; Pajdla, T. Robust wide-baseline stereo from maximally stable extremal regions. *Image Vis. Comput.* **2004**, *22*, 761–767. [CrossRef]
57. Shi, C.; Wang, C.; Xiao, B.; Zhang, Y.; Gao, S. Scene text detection using graph model built upon maximally stable extremal regions. *Pattern Recogn. Lett.* **2013**, *34*, 107–116. [CrossRef]
58. Gomez, L.; Karatzas, D. MSER-Based Real-Time Text Detection and Tracking. In Proceedings of the 2014 22nd International Conference on Pattern Recognition, Washington, DC, USA, 24–28 August 2014; pp. 3110–3115.
59. Feng, Y.; Song, Y.; Zhang, Y. Scene text localization using extremal regions and Corner-HOG feature. In Proceedings of the 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO), Zhuhai, China, 6–9 December 2015; pp. 881–886.
60. Zhang, X.; Gao, X.; Tian, C. Text detection in natural scene images based on color prior guided MSER. *Neurocomputing* **2018**, *307*, 61–71. [CrossRef]
61. Agrahari, A.; Ghosh, R. Multi-Oriented Text Detection in Natural Scene Images Based on the Intersection of MSER With the Locally Binarized Image. *Procedia Comput. Sci.* **2020**, *171*, 322–330. [CrossRef]
62. Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Structured Output Learning for Unconstrained Text Recognition. *arXiv* **2014**, arXiv:1412.5903.
63. 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 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; Volume 1801, pp. 1–10.

64. Liu, Y.; Chen, H.; Shen, C.; He, T.; Jin, L.; Wang, L. ABCNet: Real-time Scene Text Spotting with Adaptive Bezier-Curve Network**YL and HC contributed equally to this work. This work was done when Yuliang Liu was visiting The University of Adelaide. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–18 June 2020; pp. 9806–9815.
65. 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 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7553–7563.
66. Liao, M.; Wan, Z.; Yao, C.; Chen, K.; Bai, X. Real-Time Scene Text Detection with Differentiable Binarization. *AAAI* **2020**, *34*, 11474–11481. [CrossRef]
67. 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 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; Volume 1704, pp. 2642–2651.
68. 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.
69. He, T.; Tian, Z.; Huang, W.; Shen, C.; Qiao, Y.; Sun, C. An End-to-End TextSpotter with Explicit Alignment and Attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5020–5029.
70. Kim, K.-H.; Hong, S.; Roh, B.; Cheon, Y.; Park, M. PVANET: Deep but Lightweight Neural Networks for Real-time Object Detection. *arXiv* **2016**, arXiv:1608.08021.
71. Deng, D.; Liu, H.; Li, X.; Cai, D. PixelLink: Detecting Scene Text via Instance Segmentation. *arXiv* **2018**, arXiv:1801.01315.
72. Xu, Y.; Wang, Y.; Zhou, W.; Wang, Y.; Yang, Z.; Bai, X. TextField: Learning a Deep Direction Field for Irregular Scene Text Detection. *IEEE Trans. Image Proc.* **2019**, *28*, 5566–5579. [CrossRef]
73. Li, Y.; Wu, Z.; Zhao, S.; Wu, X.; Kuang, Y.; Yan, Y.; Ge, S.; Wang, K.; Fan, W.; Chen, X.; et al. PSENet: Psoriasis Severity Evaluation Network. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 800–807.
74. Tian, Z.; Huang, W.; He, T.; He, P.; Qiao, Y. Detecting Text in Natural Image with Connectionist Text Proposal Network. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Cham, Switzerland, 2016; pp. 56–72.
75. 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.
76. Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-Oriented Scene Text Detection via Rotation Proposals. *arXiv* **2017**, arXiv:1703.01086. [CrossRef]
77. Lyu, P.; Liao, M.; Yao, C.; Wu, W.; Bai, X. Mask TextSpotter: An End-to-End Trainable Neural Network for Spotting Text with Arbitrary Shapes. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
78. Liu, Y.; Zhang, S.; Jin, L.; Xie, L.; Wu, Y.; Wang, Z. Omnidirectional Scene Text Detection with Sequential-free Box Discretization. *arXiv* **2019**, arXiv:1906.02371.
79. Shi, B.; Bai, X.; Belongie, S. Detecting Oriented Text in Natural Images by Linking Segments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3482–3490.
80. He, P.; Huang, W.; He, T.; Zhu, Q.; Qiao, Y.; Li, X. Single Shot Text Detector with Regional Attention. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3066–3074.
81. Liao, M.; Shi, B.; Bai, X. TextBoxes++—A Single-Shot Oriented Scene Text Detector. *IEEE Trans. Image Proc.* **2018**, *27*, 3676–3690. [CrossRef] [PubMed]
82. Liao, M.; Zhu, Z.; Shi, B.; Xia, G.-S.; Bai, X. Rotation-Sensitive Regression for Oriented Scene Text Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; Volume 1803, pp. 5909–5918.
83. Long, S.; Ruan, J.; Zhang, W.; He, X.; Wu, W.; Yao, C. *Computer Vision—ECCV 2018, 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part II*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 19–35.

84. Baek, Y.; Lee, B.; Han, D.; Yun, S.; Lee, H. Character Region Awareness for Text Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
85. Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Reading Text in the Wild with Convolutional Neural Networks. *Int. J. Comput. Vis.* **2015**, *116*, 1–20. [CrossRef]
86. 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.
87. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Las Condes, CL, USA, 11–18 December 2015; pp. 1440–1448.
88. Xiang, D.; Guo, Q.; Xia, Y. Robust Text Detection with Vertically-Regressed Proposal Network. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Cham, Switzerland, 2016; pp. 351–363.
89. Liao, M.; Shi, B.; Bai, X.; Wang, X.; Liu, W. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. *arXiv* **2016**, arXiv:1611.06779.
90. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37.
91. Liu, Y.; Jin, L. Deep Matching Prior Network: Toward Tighter Multi-oriented Text Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3454–3461.
92. Wang, S.; Liu, Y.; He, Z.; Wang, Y.; Tang, Z. A quadrilateral scene text detector with two-stage network architecture. *Pattern Recogn.* **2020**, *102*, 107230. [CrossRef]
93. 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]
94. Deng, L.; Gong, Y.; Lu, X.; Lin, Y.; Ma, Z.; Xie, M. STELA: A Real-Time Scene Text Detector with Learned Anchor. *IEEE Access* **2019**, *7*, 153400–153407. [CrossRef]
95. Cai, Y.; Wang, W.; Ren, H.; Lu, K. SPN: Short path network for scene text detection. *Neural Comput. Appl.* **2019**, *32*, 6075–6087. [CrossRef]
96. Xue, C.; Lu, S.; Zhang, W. MSR: Multi-Scale Shape Regression for Scene Text Detection. *arXiv* **2019**, arXiv:1901.02596.
97. Wang, Y.; Xie, H.; Zha, Z.; Xing, M.; Fu, Z.; Zhang, Y. ContourNet: Taking a Further Step toward Accurate Arbitrary-shaped Scene Text Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–18 June 2020.
98. Wang, Y.; Xie, H.; Fu, Z.; Zhang, Y. DSRN: A Deep Scale Relationship Network for Scene Text Detection. *IJCAI* **2019**, 947–953. [CrossRef]
99. Zhang, C.; Liang, B.; Huang, Z.; En, M.; Han, J.; Ding, E.; Ding, X. Look More Than Once: An Accurate Detector for Text of Arbitrary Shapes. *arXiv* **2019**, arXiv:1904.06535.
100. Xing, L.; Tian, Z.; Huang, W.; Scott, M.R. Convolutional Character Networks. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 2 November 2019; Volume 1910.
101. Zhang, S.; Liu, Y.; Jin, L.; Wei, Z.; Shen, C. OPMP: An Omni-directional Pyramid Mask Proposal Network for Arbitrary-shape Scene Text Detection. *IEEE Trans. Multimed.* **2020**. [CrossRef]
102. 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.
103. Li, X.; Wang, W.; Hou, W.; Liu, R.-Z.; Lu, T.; Yang, J. Shape Robust Text Detection with Progressive Scale Expansion Network. *arXiv* **2018**, arXiv:1806.02559.
104. Tang, J.; Yang, Z.; Wang, Y.; Zheng, Q.; Xu, Y.; Bai, X. SegLink++: Detecting Dense and Arbitrary-shaped Scene Text by Instance-aware Component Grouping. *Pattern Recogn* **2019**, *96*, 106954. [CrossRef]
105. 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), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 764–772. [CrossRef]

106. Zhou, Z.; Wu, S.; Kong, S.; Zheng, Y.; Ye, H.; Chen, L.; Pu, J. Curve Text Detection with Local Segmentation Network and Curve Connection. *arXiv* **2019**, arXiv:1903.09837.
107. Wang, W.; Xie, E.; Song, X.; Zang, Y.; Wang, W.; Lu, T.; Yu, G.; Shen, C. Efficient and Accurate Arbitrary-Shaped Text Detection with Pixel Aggregation Network. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019.
108. 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, Australia, 20–25 September 2019; pp. 375–380. [CrossRef]
109. Kobchaisawat, T.; Chalidabhongse, T.H.; Satoh, S. Scene Text Detection with Polygon Offsetting and Border Augmentation. *Electronics* **2020**, *9*, 117. [CrossRef]
110. Wang, C.; Yin, F.; Liu, C.-L. Scene Text Detection with Novel Superpixel Based Character Candidate Extraction. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; pp. 929–934. [CrossRef]
111. Xue, C.; Lu, S.; Zhan, F. Accurate Scene Text Detection through Border Semantics Awareness and Bootstrapping. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Springer: Cham, Switzerland, 2018.
112. Tian, Z.; Shu, M.; Lyu, P.; Li, R.; Zhou, C.; Shen, X.; Jia, J. Learning Shape-Aware Embedding for Scene Text Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 4229–4238.
113. Zhong, Z.; Sun, L.; Huo, Q. An anchor-free region proposal network for Faster R-CNN-based text detection approaches. *Int. J. Doc. Anal. Recognit. Ijdar.* **2018**, *22*, 315–327. [CrossRef]
114. Wang, Q.; Zheng, Y.; Betke, M. SA-Text: Simple but Accurate Detector for Text of Arbitrary Shapes. *arXiv.* **2019**, arXiv:1911.07046.
115. Wu, W.; Xing, J.; Zhou, H. TextCohesion: Detecting Text for Arbitrary Shapes. *arXiv* **2019**, arXiv:1904.12640.
116. Liu, Z.; Lin, G.; Yang, S.; Liu, F.; Lin, W.; Goh, W.L. Towards Robust Curve Text Detection with Conditional Spatial Expansion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 7261–7270.
117. Lucas, S.M.; 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, Edinburgh, UK, 6 August 2003; pp. 682–687.
118. Lee, S.; Cho, M.; Jung, K.; Kim, J.H. Scene Text Extraction with Edge Constraint and Text Collinearity. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 3983–3986.
119. Shahab, A.; Shafait, F.; Dengel, A. ICDAR 2011 Robust Reading Competition Challenge 2: Reading Text in Scene Images. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 18–21 September 2011; Volume 1, pp. 1491–1496.
120. Yao, C.; Bai, X.; Liu, W.; Ma, Y.; Tu, Z. Detecting texts of arbitrary orientations in natural images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, Rhode Island, 16–21 June 2012; Volume 1, pp. 1083–1090.
121. Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; Bigorda, L.G.i.; Mestre, S.R.; Mas, J.; Mota, D.F.; Almazan, J.A.; De Las Heras, L.P. 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]
122. Yin, X.-C.; Yin, X.; Huang, K.; Hao, H.-W. Robust Text Detection in Natural Scene Images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 970–983.
123. Veit, A.; Matera, T.; Neumann, L.; Matas, J.; Belongie, S. COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images. *arXiv* **2016**, arXiv:1601.07140.
124. Gupta, A.; Vedaldi, A.; Zisserman, A. Synthetic Data for Text Localisation in Natural Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2315–2324.

125. Yuan, T.-L.; Zhu, Z.; Xu, K.; Li, C.-J.; Hu, S.-M. Chinese Text in the Wild. *arXiv* **2018**, arXiv:1803.00085.
126. Shi, B.; Yao, C.; Liao, M.; Yang, M.; Xu, P.; Cui, L.; Belongie, S.; Lu, S.; Bai, X. ICDAR2017 Competition on Reading Chinese Text in the Wild (RCTW-17). In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; pp. 1429–1434. [CrossRef]
127. Ch'ng, C.K.C.; Chan, C.S. Total-Text: A Comprehensive Dataset for Scene Text Detection and Recognition. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; pp. 935–942. [CrossRef]
128. Liu, Y.; Jin, L.; Zhang, S.; Luo, C.; Zhang, S. Curved Scene Text Detection via Transverse and Longitudinal Sequence Connection. *Pattern Recogn.* **2019**, *90*, 337–345. [CrossRef]
129. 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), Kyoto, Japan, 9–15 November 2017; pp. 1454–1459. [CrossRef]
130. Chng, C.-K.; Liu, Y.; Sun, Y.; Ng, C.C.; Luo, C.; Ni, Z.; Fang, C.; Zhang, S.; Han, J.; Ding, E.; et al. ICDAR2019 Robust Reading Challenge on Arbitrary-Shaped Text (RRC-ArT). In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 22–25 September 2019.
131. Nayef, N.; Patel, Y.; Busta, M.; Chowdhury, P.N.; Karatzas, D.; Khlif, W.; Matas, J.; Pal, U.; Burie, J.-C.; Liu, C.; et al. ICDAR2019 Robust Reading Challenge on Multi-lingual Scene Text Detection and Recognition—RRC-MLT-2019. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 22–25 September 2019.
132. Sun, Y.; Liu, J.; Liu, W.; Han, J.; Ding, E.; Liu, J. Chinese Street View Text: Large-scale Chinese Text Reading with Partially Supervised Learning. In Proceedings of the IEEE International Conference on Computer Vision; Seoul, Korea, 27 October–2 November 2020; pp. 9085–9094.
133. Xie, E.; Zang, Y.; Shao, S.; Yu, G.; Yao, C.; Li, G. Scene Text Detection with Supervised Pyramid Context Network. *arXiv* **2018**, arXiv:1811.08605. [CrossRef]
134. 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 IEEE International Conference on Computer Vision, Venice, Italy, 29 October 2017; pp. 4950–4959.
135. Xing, D.; Li, Z.; Chen, X.; Fang, Y. ArbiText: Arbitrary-Oriented Text Detection in Unconstrained Scene. *arXiv* **2017**, arXiv:1711.11249.
136. Zhu, X.; Jiang, Y.; Yang, S.; Wang, X.; Li, W.; Fu, P.; Wang, H.; Luo, Z. Deep Residual Text Detection Network for Scene Text. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; pp. 807–812. [CrossRef]
137. Liu, Z.; Lin, G.; Yang, S.; Feng, J.; Lin, W.; Goh, W.L. Learning Markov Clustering Networks for Scene Text Detection. *arXiv* **2018**, arXiv:1805.08365.
138. 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]
139. Roy, S.; Shivakumara, P.; Pal, U.; Lu, T.; Kumar, G.H. Delaunay triangulation based text detection from multi-view images of natural scene. *Pattern Recogn. Lett.* **2020**, *129*, 92–100. [CrossRef]
140. Wang, P.; Zhang, C.; Qi, F.; Huang, Z.; En, M.; Han, J.; Liu, J. A Single-Shot Arbitrarily-Shaped Text Detector based on Context Attended Multi-Task Learning. In Proceedings of the 27th ACM International Conference on Multimedia, New York, NY, USA, 21–25 October 2019; pp. 1277–1285. [CrossRef]
141. Zhu, Y.; Du, J. Sliding Line Point Regression for Shape Robust Scene Text Detection. *arXiv* **2018**, arXiv:1801.09969.
142. Kang, L.; Li, Y.; Doermann, D. Orientation Robust Text Line Detection in Natural Images. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 4034–4041. [CrossRef]
143. Zhang, L.; Liu, Y.; Xiao, H.; Yang, L.; Zhu, G.; Shah, S.A.; Bennamoun, M.; Shen, P. Efficient Scene Text Detection with Textual Attention Tower. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020.

144. Liu, H.; Guo, A.; Jiang, D.; Hu, Y.; Ren, B. PuzzleNet: Scene Text Detection by Segment Context Graph Learning. *arXiv* **2020**, arXiv:2002.11371.
145. Dasgupta, K.; Das, S.; Bhattacharya, U. Scale-Invariant Multi-Oriented Text Detection in Wild Scene Images. *arXiv* **2020**, arXiv:2002.06423.