applied sciences

MDPI

Article

Scene Text Detection Using Attention with Depthwise Separable Convolutions

Ehtesham Hassan * and Lekshmi V. L.

Department of Computer Science and Engineering, Kuwait College of Science and Technology, Doha 35004, Kuwait; l.vijayan@kcst.edu.kw
* Correspondence: e.hassan@kcst.edu.kw; Tel.: +965-24972865

Abstract: In spite of significant research efforts, the existing scene text detection methods fall short of the challenges and requirements posed in real-life applications. In natural scenes, text segments exhibit a wide range of shape complexities, scale, and font property variations, and they appear mostly incidental. Furthermore, the computational requirement of the detector is an important factor for real-time operation. To address the aforementioned issues, the paper presents a novel scene text detector using a deep convolutional network which efficiently detects arbitrary oriented and complex-shaped text segments from natural scenes and predicts quadrilateral bounding boxes around text segments. The proposed network is designed in a U-shape architecture with the careful incorporation of skip connections to capture complex text attributes at multiple scales. For addressing the computational requirement of the input processing, the proposed scene text detector uses the MobileNet model as the backbone that is designed on depthwise separable convolutions. The network design is integrated with text attention blocks to enhance the learning ability of our detector, where the attention blocks are based on efficient channel attention. The network is trained in a multi-objective formulation supported by a novel text-aware non-maximal procedure to generate final text bounding box predictions. On extensive evaluations on *ICDAR2013*, *ICDAR2015*, *MSRA-TD500*, and *COCOText* datasets, the paper reports detection F-scores of `0.910`, `0.879`, `0.830`, and `0.617`, respectively.

Keywords: scene text detection; MobileNets; convolutional network; text attention

Citation: Hassan, E.; L., L.V. Scene Text Detection Using Attention with Depthwise Separable Convolutions. *Appl. Sci.* **2022**, *12*, 6425. https://doi.org/10.3390/app12136425

Academic Editor: Andrea Prati

Received: 9 May 2022
Accepted: 20 June 2022
Published: 24 June 2022

1. Introduction

Text processing in images and videos is an important problem in many digital applications. With advancements in mobile and augmented reality applications, efficient scene text processing can significantly improve the overall experience in such applications. Unlike conventional optical character recognition, scene text processing applications deal with much larger variations in text appearances in terms of the orientations, shapes, scripts, and scales. The localization of text segments in input images forms the precursor for understanding the texts embedded in scene images. In this case, a detector localizes the text regions in the input images as rectangular or quadrilateral bounding boxes.

Early research on scene text detection focused on developing handcrafted features for modeling text components in natural scenes [1–4]. These methods exploited the conventional image processing tools including edge detection, connected component analysis, morphological operations, and neighborhood analysis for content modeling in images and videos [5–7]. Stroke width transform (SWT) by Epshtein et al. [2] and maximally stable extremal regions (MSERs) by Neumann et al. [8] are two important methods in this body of research. The SWT presented steps for pixel-level character stroke width analysis for text instance segmentation in images, whereas the MSER for scene text detection applied contiguous image segment analysis at multiple levels. The concepts proposed in SWT and MSER were subsequently in many methods for scene text analysis [9–12]. However, these methods fail to address the variations and complexities in real-life scene text detection due to the inherently parameter-sensitive design steps.

With the success of deep neural networks-based image recognition [13], many recent methods have explored various designs of deep convolutional neural networks for scene text analysis [14,15]. The majority of these methods focused on improving text detection performance by applying complex convolutional neural net architectures to enhance the representation power of detection models. However, this limits the applicability of developed methods in real-time applications because of computational requirements. In addition, despite significant effort, the existing convolutional architectures fail to handle complex scene text appearances in natural scenes as observed from the state of the art on public datasets [16–18].

To this end, the paper presents a novel scene text detector based on convolutional neural networks for predicting quadrilateral bounding boxes on scene text instances. In addition to the detection accuracy, our approach focuses on a computationally lightweight and accurate scene text detector suitable for digital applications with limited computational resources, e.g., mobile and embedded applications. The proposed detector uses the MobileNet model as the backbone, where the detector network is designed in a U-shaped encoder–decoder architecture by introducing skip connections for regulated feature fusion at multiple levels. The selection of MobileNet was made for its lightweight and efficient design based on the depthwise separable convolution operation [19].

The challenges of text detection in natural scenes bridge over to conventional object detection and object instance segmentation in images. Therefore, the proposed detector is trained in multitask learning, combining the text instance segmentation with quadrilateral bounding box prediction around text instances. The combination of dual objectives guides the detector to capture text attributes for arbitrarily oriented, complex shaped, and styled scene text detection. In this convention, the subsequent layers in sequential convolutional network design extract finer details from the input feature map by information aggregation using layerwise convolutional operation and feature sampling. In this respect, many recent works on object detection have explored the visual attention mechanism in the deep network design for guiding the network to focus on target specific details [20–23]. Along the same lines, our approach incorporates carefully placed text attention blocks in the detector network design for enhancing the representative power of extracted features. The common design of attention mechanisms includes convolutional blocks, which increase the overall model complexity and computational requirements. The text attention blocks in the proposed detector are based on efficient channel attention (ECA) [24], which has demonstrated excellent performance on object detection tasks with marginal computational cost. The following are the major contributions in this paper.

- A novel U-shaped network for scene text detection for real-time applications is presented, which applies the depthwise separable convolution operation. The network predicts a quadrilateral bounding box around text instances in scene images. The incorporation of skip connections enables the detector to efficiently capture arbitrarily shaped text instances at multiple scales. The network is incorporated with ECA-based text attention blocks for robust and efficient text feature extraction, and the training is performed using a novel multitasking formulation. Next, a novel post-processing method using non-maximal suppression is applied for final prediction, which accounts for the text expectation in candidate quadrilateral bounding boxes.
- The different components of the proposed detector have been extensively validated on *ICDAR2013*, *ICDAR2015*, *COCOText*, and *MSRA-TD500* datasets. With thorough experiments under different settings, the results demonstrate that the proposed scene text detector presents an efficient solution for the detection of arbitrary shaped, multi-oriented text instances in different real-life settings. As shown later in the results, the proposed detector outperforms many prominent deep neural network-based methods, and it achieves on par performance in comparison with others.

The paper structure is as follows. Section 2 presents relevant works pertaining to the deep learning methods for scene text detection. The proposed scene text detector is discussed in Section 3. The experimental validation of the proposed methodology and

relevant discussions are available in Sections 4 and 5. The final section summarizes our contribution and discusses the direction of future work.

2. Literature Survey

Delakis et al. [25] presented an early application of convolutional networks designed with convolution, sampling, and activation layers for text detection in images. Subsequently, Wang et al. [26] demonstrated a convolutional network for scene text recognition combining text localization, character recognition, and lexicon-driven word recognition. On similar lines, Jaderberg et al. [27] demonstrated jointly trained convolutional networks with shared weights for text detection and recognition. However, these methods analyze the input by spatial scanning in the image space, which is computationally inefficient.

With progress in convolutional neural network based object detectors, e.g., Fast RCNN [28], RFCN [29], SSD [30], and Mask RCNN [31], many recent scene text detection methods followed similar designs to directly predict boundary boxes around the text segments [32–37]. The early deep convolutional neural network design for text detection focused on axis-aligned rectangular box prediction around text segments [38–42]. The method by Tian et al. [40] presented a convolution RNN for text line proposal generation and merging. Similarly, [39] showed the application of residual features in a convolutional RNN framework for scene text detection. However, the textual content in natural images is never constrained within the rectangular boundary but appears in arbitrary orientations and curved shapes within complex backgrounds. Wang et al. [43] proposed a text proposal network founded on the RPN design in the Faster RCNN object detector. The proposals are refined using an LSTM to generate polygonal bounding boxes around text segments. In [33], Zhang et al. designed a fully convolutional network for detecting the text saliency map in a given input, which is subsequently processed by the MSER for text detection. CharNet [44] presented a convolutional architecture combining the character and text instance detection in a single network. There are also methods that focus on the segmentation of text regions for the detection of irregular and complex text instances [34,45,46]. Mask Textspotter [34] applied the Fast RCNN detector with a convolutional branch for text and character-level image segmentation. PixelLink [46] presented a novel formulation for modeling image-level pixel associations using a convolutional architecture. Xie et al. [47] and Huang et al. [48] presented deep convolutional neural architectures by processing Mask RCNN generated text instances combined with feature pyramids.

The direct regression of bounding boxes around text instances has been another research approach to solve scene text detection. Regarding the direct regression for multi-oriented scene texts, He et al. [49], EAST [32], SegLink [50], and Lyu et al. [51] presented some prominent methods based on deep learning for regression. EAST presented a U-shaped convolutional architecture for the prediction of words or text lines, whereas SegLink presented a convolutional architecture for learning associations between text segments at multiple levels of feature maps. Deng et al. [52] proposed a convolutional architecture network to regress the corner points around text segments, which are combined to generate quadrilateral text boxes. The LOMO detector [53] again pursued the regression of corner points around text segments with additional convolutional modules for the refinement of bounding box prediction and text shape learning. In addition, there are methods for solving the problem in a bottom–up manner following the character/text symbol detection followed by sequence analysis for word and text line detection [33,42,44,54]. However, the majority of these methods have ignored the computational requirement for the proposed method, which is critical in many applications such as mobile and augmented reality. There are also some works on scene text detection for applications with limited computational support [55–58]. However, these methods fail to address significant variations in the color, scale, orientation, aspect ratio, and shape of text instances in real-life scenes.

To address the issues mentioned above, our method presents a novel convolutional architecture to predict quadrilateral bounding boxes around text segments of different shapes and orientations in natural scenes. The proposed model is based on the MobileNet

model that uses the depthwise separable convolutional operation, resulting in more computationally efficient object detectors than the convolutional network based detectors. There exists an earlier application of MobileNet for scene text detection in [59], although it was limited in exploration and experiments. We present a novel U-shaped convolutional neural architecture for scene text detection using MobileNet incorporating text attention blocks in the feature extraction stage. The text attention in the proposed network is based on the ECA module [24], which is an efficient method for implementing channel attention in convolutional networks. In contrast to other methods, such as [32,33,39,40,50], our method uses attention blocks in the convolutional architecture to improve feature extraction on text-specific attributes. The detector is trained using a novel multitasking formulation based on focal loss, which also accounts for the skewed sample distribution in scene text problems. The predictions by the detector network are processed by a novel non-maximal suppression technique to account for text expectation in predicted quadrilateral bounding boxes.

3. Proposed Methodology

The following discussion presents the proposed scene text detector for quadrilateral bounding box predictions around text segments. First, the proposed network architecture is introduced, which is followed by the training objectives and post-processing steps.

3.1. The Network Design—*MobileTDNet*

The proposed detector (**MobileTDNet**) is designed to accurately detect scene text segments in complex shapes, different scales and styles, and multiple orientations. In addition, the detector network should be computationally lightweight and efficient. The MobileNet architecture-based convolutional neural models use optimized convolution operations in the network design [19], where the layer-wise convolution operations are separated as multiplication and addition operation on the depth dimension of feature maps. MobileNets have been shown to be highly effective and efficient in many vision problems. Therefore, the proposed scene detector is designed on the MobileNet model as the backbone. Figure 1 presents the illustration of the proposed scene text detector network. The network consists of three major branches: feature extraction, feature merging, and output layers.

The feature extraction branch generates convolutional feature maps of multiple spatial resolutions. The branch is designed by the stacking of convolutional blocks, as shown in Figure 1. Each convolutional block consists of a convolution layer with a 3×3 mask followed by another convolution layer with a 1×1 mask. The first convolution layer in the block is designed with depthwise separable convolution operation. The subsequent convolution layer with a 1×1 mask reduces the number of channels in the generated feature map. The MobileNet base is extended by drawing out convolutional features at multiple scales. These intermediate features capture text shape and scale complexities with varying degrees of detail. The following five intermediate feature maps—$F^{\frac{1}{2}}$, $F^{\frac{1}{4}}$, $F^{\frac{1}{8}}$, $F^{\frac{1}{16}}$ and $F^{\frac{1}{32}}$ in **MobileTDNet** are drawn out at $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$ and $\frac{1}{32}$th scale of the input resolution. The last fully connected layer from the MobileNet base is removed, and we subsequently process the feature map F^{OUT} with intermediate feature maps extracted above in the hierarchy. For strengthening the extracted features, text attention blocks are incorporated at three steps in the feature extraction branch, as illustrated in Figure 1. The details of attention blocks are discussed later in Section 3.3.

In the feature merging branch, the F^{OUT} is recursively combined with $F^i \in \{F^{\frac{1}{2}}, F^{\frac{1}{4}}, F^{\frac{1}{8}}, F^{\frac{1}{16}}, F^{\frac{1}{32}}\}$ through the skip connections. A skip connection is defined by two inputs, the intermediate feature map F^i and the output from the Upscale block, which processes the skip connection output from the previous level. Figure 1 shows the Upscale block design consisting of a two-dimensional upsample layer based on bilinear interpolation, which is followed by a convolutional layer for data smoothening. The channel dimension of the output feature map is reduced with a convolution filter with a 1×1 mask to align

with the second merge input to skip connections. The convolutional layer outputs in the detector are processed through batch normalization and ReLU activation.

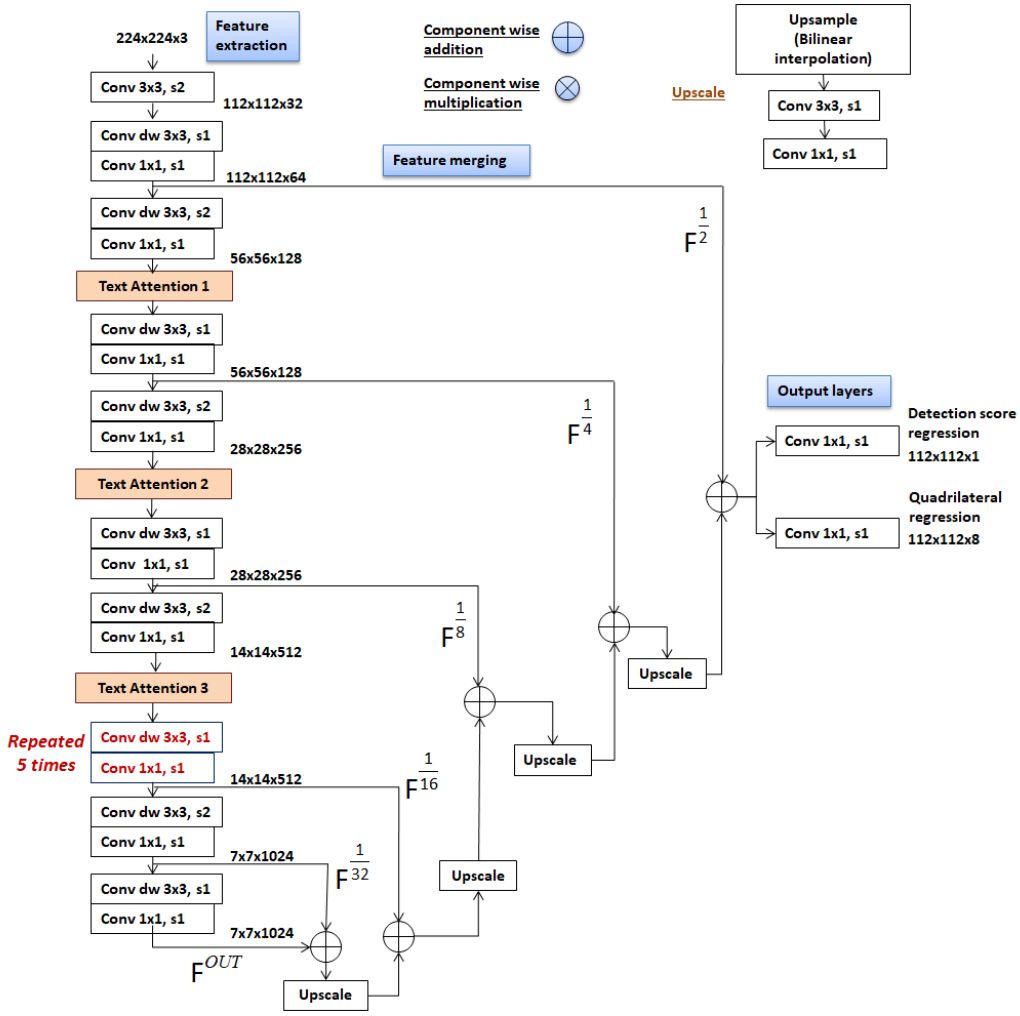


Figure 1. MobileTDNet scene text detection architecture.

The network is designed with two output layers generating the following predictions at one-half of the input resolution: (i) text/non-text semantic map, and (ii) text bounding box regression at each position. The network predictions are processed through a novel text aware non-maximal suppression method to generate the final prediction. The output layers are described below.

- First output layer: 2nd order tensor (out_1) of size $\frac{S_1}{2} \times \frac{S_2}{2}$ having single channel output, assuming the input size as $S_1 \times S_2$. The (out_1) corresponds to the text/non-text segmentation map generated at one-half the resolution of the input. The tensor values represent text confidence scores at each position.
- Second output layer: 3rd order tensor (out_2) of size $\frac{S_1}{2} \times \frac{S_2}{2} \times 8$ encodes the pixel-level quadrilateral bounding box predictions at the output feature map. On the depth of out_2, each pair of values corresponds to a corner of the predicted quadrilateral.

3.2. The Loss Function

The **MobileTDNet** is trained in multitask learning accounting the predictions from both output layers. The training loss functions is defined as follows:

$$L = L_{seg} + \lambda_r L_{reg} \tag{1}$$

The L_{seg} component in Equation (1) represents the segmentation loss, measuring the network's ability to identify text and non-text regions in the input. Simultaneously, the loss component L_{reg} measures the network's ability to correctly predict the geometric position of text bounding boxes around text lines in the input. Here, λ_r represents the regularization parameter. Throughout all the experiments in this work, the parameter λ_r is set equal to 1, i.e., both loss components contribute equally.

The cross-entropy loss, commonly applied in semantic image segmentation tasks, is used for evaluating the segmentation loss L_{seg}. The out_1 values are used for calculating L_{seg}. In this case, the loss function measures the difference between the predicted probability distribution and the actual distribution. The conventional cross-entropy loss does not consider the difference between the densities of relevant and non-relevant samples. However, the text/non-text class imbalance is a practical challenge in scene text detection, which limits the detector training performance. To address the issue, online hard example mining [60] has been a popular method for maintaining the balance between positive and negative classes. However, the method increases the number of computational steps and memory requirements in the network training process. In the present formulation, the focal loss function by Lin et al. [61] is applied to measure the segmentation error, which accounts for the text and non-text sample distribution in scene detection. The focal loss incorporates an additional weight factor in the conventional cross-entropy function, which balances the positive and negative classes.

$$\begin{aligned} L_{seg}(c_t) = &- \alpha y^* (1 - c_t)^{\gamma} log(c_t) \\ &- (1 - \alpha)(1 - y^*) c_t^{\gamma} log(1 - c_t) \end{aligned} \tag{2}$$

The focal loss expression for output c_t with y^*as the ground truth is represented by Equation (2). The term $(1 - c_t)^{\gamma}$ contributes as the density modulating factor in the loss function, with γ as the tunable parameter. As observed, the modulating factor scales down the contribution of easy samples in L_{seg} based on the selected value of γ. Following the analysis presented in [61], the parameter γ is set equal to 2 to conduct all experiments discussed in this work. Furthermore, the expression also incorporates the parameter α as a density-based balancing factor between two classes, which is computed as follows.

$$\alpha = 1 - \frac{\text{\#of text pixels in ground truth}}{\text{\#of pixels in ground truth}} \tag{3}$$

The component L_{reg} accounts for the quadrilateral text bounding box prediction error. The ith position in out_2 represents corners of the predicted quadrilateral bounding box in the ordered set

$$q_i = \{(h_j, w_j), \cdots | j \in 0, 1, 2, 3\}$$

The position loss L_{reg} refers to the geometric error in bounding box coordinate prediction with respect to the corresponding ground truth $\hat{q}_i$. The loss is calculated using the smooth $L1$ loss function as follows

$$L_{reg}(\hat{q}_i, q_i) = \sum_{\hat{h}_j \in \hat{q}_i, h_j \in q_i} smooth_{L_1}(\hat{h}_j - h_j) \tag{4}$$

$$\text{where, } smooth_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1, \\ |x - 0.5| & \text{otherwise} \end{cases}$$

3.3. Design of Text Attention Blocks

Attention mechanisms have been successfully applied to improve the effectiveness of convolutional neural network architectures along with the careful selection of layers, filters, and channel dimensions. As shown in Figure 1, the proposed detector includes text attention blocks at three stages of the feature extraction. The objective of the attention block at the early stage of feature extraction is to channelize the feature extraction process

toward capturing the text positional information. The subsequent attention blocks enables the feature extraction process to capture text attributes at all scales. The text attention blocks in the proposed detector are implemented using the efficient channel attention (ECA) mechanism by Wang et al. [24]. The ECA mechanism exploits the local cross-channel attention implemented through one-dimensional convolutional operators with an adaptive kernel size. The kernel size k represents the neighbors to be considered for the cross-channel interaction exploration. Figure 2 shows the basic design of efficient channel attention blocks adopted from [24]. The channel dimension Ch and the kernel size k are related by the nonlinear mapping $Ch = 2^{\gamma * k - b}$ i.e.,

$$k = \left| \frac{log_2(Ch) + b}{\gamma} \right|_{\text{nearest odd value}} \tag{5}$$

The implementation of the ECA module includes a channel-wise feature aggregation layer using global average pooling followed by processing through a one-dimensional convolutional layer. The convolutional layer output is passed through a fully connected layer that outputs the weight of different channels. The **MobileTDNet** network uses a similar block structure. The input feature aggregation is an important step in the ECA mechanism. The ECA-Net in [24] builds on global average pooling for channel-wise feature aggregation following SENet [62]. The average pooling-based feature values correlate with the target shape extent in the given input, whereas the max pooling-based feature values correlate with detection object-specific attributes, which are effective in the case of target objects appearing at different scales.

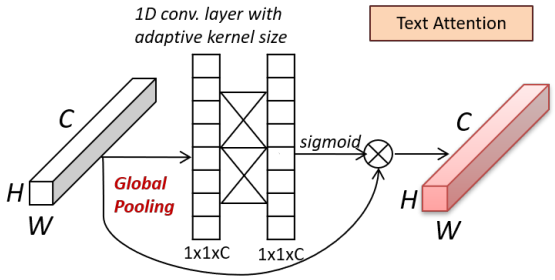


Figure 2. Design of the Text Attention Block.

Therefore, the different feature aggregation strategies based attention blocks are applied. At the early stage of feature processing in the network, the **Text Attention 1** block is implemented with global average pooling-based feature aggregation to complement the text discovery process in the input. In later stages where the network generated feature maps capture the finer text segment details, the **Text Attention 2** and **Text Attention 3** attention blocks are implemented with the global max pooling-based feature aggregation, since this affects the detection of small text segments. The parameters γ and b in Equation (5) control the span of the channel neighborhood for attention analysis. Without loss of generality, the parameter b and γ are set to 2 and 3, respectively, for all experiments in this work.

3.4. Text Aware Non-Maximal Suppression

The **MobileTDNet** network generates the following outputs:

- Text/non-text segmentation map $C = \{c_1, \cdots, c_n\}$ with c_i representing the text confidence score at the ith position with $n = |\frac{S_1}{2} \times \frac{S_2}{2}|$.
- Quadrilateral predictions $Q = \{q_1, \cdots, q_n\}$ where q_i represents quadrilateral bounding box prediction at the ith position.

For generating the final quadrilateral text bounding box predictions, the non-maximal suppression (NMS) technique [63] is applied in the following steps:

1. To identify the dominant predictions corresponding to the text segments, the text/non-text segmentation map C is filtered at 0.5. The filtering results in a reduced set of

confidence score predictions C_t, and overlapping text bounding box predictions Q_t. The NMS technique reduces the set of candidate bounding boxes by analyzing the confidence score associated with bounding box prediction and pairwise overlap between candidates. In contrast, the **MobileTDNet** output layer 2 prediction consists of only the geometric position of text bounding boxes without the measure of text attributes within the bounding box.

2. For calculating the text attribute of a quadrilateral box prediction q_i, the text neighborhood correlation within the quadrilateral boundary is exploited. For each quadrilateral prediction q_i, the text confidence scores from C_t at all positions within the boundary of q_i is averaged to a single point measure c_i^q. This returns the set $C_Q = \{c_i^q, \cdots | i = 0 \text{ to } |Q_t|\}$. Here, c_i^q is referred to as the text expectation measure for the bounding box q_i, as text appearances within the boundary can be in any shape and size, and the constricted boundary around the text segment is expected to have high text expectation. The quadrilateral boxes with low text expectation measures are filtered out, and the boxes within the top 10% of the text expectation measures are preserved. The filtration results in the quadrilateral bounding box prediction set Q_F, with the corresponding text expectation values represented in set C_F. The Q_F and C_F sets are next processed using the NMS to generate final predictions.
3. The conventional NMS procedure outcomes are sensitive to the selected intersection over union (IoU) threshold; therefore, the soft-NMS procedure by Bodla et al. [64] is applied. The soft-NMS uses a fixed IoU threshold for pruning, but the confidence scores of all unfinished quadrilateral boxes are rescaled with a smooth penalty function in each pruning iteration. The idea is to gradually decrease the score of overlapping quadrilateral boxes, which is expected to reduce the contribution in the false positive rate in the detection. At the same time, the function should impose a low penalty in the case of non-overlapping quadrilateral prediction boxes. The complete algorithm steps are listed below (Algorithm 1).

Algorithm 1: Soft-NMS for final quadrilateral box prediction

Input: Quadrilateral prediction set $Q_F = \{q_1, q_2, \cdots, q_n\}$ with text estimation scores for bounding boxes as $C_F = \{c_1^q, c_2^q, \cdots, c_n^q\}$; IoU threshold iuo_{th};
Output: Final quadrilateral set $Q_{nms} = \{q_1, q_2, \cdots, q_m\}$;

while $Q_F \neq$ `NULL` **do**
  $ind = \texttt{argmax}\ C_F$;
  $q_{max} \leftarrow q_{ind} \in Q_F$;
  $c_{max} \leftarrow c_{ind} \in C_F$;
  $Q_{nms} \leftarrow Q_{nms} \cup q_{max}$;
  $Q_F \leftarrow Q_F - q_{max}$;
  $C_F \leftarrow C_F - c_{max}$;
  for i to $|Q_F|$ **do**
    if $IoU(q_i, q_{max}) \geq iou_{th}$ **then**
      $Q_F \leftarrow Q_F - q_i$;
      $C_F \leftarrow C_F - c_i$;
    end
    $c_i = c_i \texttt{exp}^{\frac{-\texttt{IoU}(q_i, q_{max})}{\texttt{sigma}}}$
  end
end

The IoU measure is computed using the originally predicted coordinates of the quadrilateral boundary. Following the analysis in [64], the `sigma` and iuo_{th} parameters are set to 0.5 and 0.3 throughout all experiments.

4. Experimental Results

For evaluation of the methodology presented in this paper, the following datasets are used.

1. *ICDAR2013* [65]: The dataset is a collection of natural images having horizontal and near-horizontal text appearances. The collection consists of 229 training and 233 testing images having character and word-level bounding box annotations and corresponding annotations.
2. *ICDAR2015* [66]: The dataset was released as the fourth challenge in the ICDAR 2015 robust reading competition (incidental scene text detection). The dataset consists of 1500 images, of which 1000 were for training purposes, and the remaining images were used for testing. The dataset images are real-life scenes captured in Google Glass in an incidental manner, with the annotations available as quadrangle text bounding boxes with corresponding unicode transcription.
3. *COCOText* [67]: This is a large dataset that consists of 63,000 images sampled from the MSCOCO image collection [68] exhibiting scene texts in all appearances. The dataset provides rich annotations, including the text bounding boxes, handwritten/printed labeling, script labeling, and transcripted text. The bounding boxes are horizontal axis aligned. The dataset comes with a standard distribution of 43,000 training images, 10,000 for validation and the remaining 10,000 for testing tasks.
4. *MSRA-TD500* [69]: The dataset consists of 500 examples distributed as 300/200 for training and testing tasks. The images are indoor and outdoor natural scenes with English and Chinese texts in all orientations and complex backgrounds. The image resolution varies between 1294×864 and 1920×1280. The annotations in the dataset are available at the text line level with the orientation value of corresponding text lines.

In all experiments, the train/test distribution, the evaluation protocol, and measures are applied as suggested in the original source. The experiments in this work focused on verification of the following attributes of the **MobileTDNet** design: (i) the networks' ability to model the text/non-text image attributes across all style variations and backgrounds, (ii) the role of skip connections in capturing the text properties at all scales; and (iii) the impact of text attention blocks in the current architecture.

4.1. Network Training and Hyperparameters

The proposed network is trained from scratch using the Adam optimizer [70,71]. Adam is an extension to the stochastic gradient descent algorithm that applies the first-order and second-order gradient moments for adapting the learning rate to network weight parameters. Table 1 shows the hyperparameters used for training **MobileTDNet** on different datasets, which were set experimentally following the protocols suggested in [71]. The **MobileTDNet** loss function intrinsically addresses the positive and negative sample unbalance; therefore, the training procedure does not include a hard negative sampling step.

Table 1. MobileTDNet hyperparameters: *lr* represents the learning rate.

Dataset	Initial *lr*	# of Epochs	Batch Size	# of Epochs for *lr* Decay
ICDAR2013	0.001	50	16	20
ICDAR2015	0.001	50	16	20
MSRA-TD500	0.0005	60	24	30
COCO-Text	0.0001	100	8	50

Pre-training: Before evaluation on different datasets, the **MobileTDNet** architecture is pre-trained on the combined training set collection of *ICDAR2013* and *ICDAR2015* datasets. The pre-training is performed for 10 epochs with a slow learning rate of 0.0001, and the batch

size is fixed at 16. The pre-training step is required as some experimental datasets are small in size, and training the network on such datasets from scratch with randomly initialized weights would not be effective. Therefore, **MobileTDNet** is pre-trained on a large and diverse collection to initialize the network weights with a high-level understanding of the task domain. Subsequently, the detector is trained on different datasets with the parameters given in Table 1. The learned lr is gradually reduced by half at every 10 epochs after crossing the initial stage of training (number of epochs mentioned in the last column of the Table 1).

Data augmentation: For training the **MobileTDNet** model on *ICDAR2013*, *ICDAR2015* and *MSRA-TD500* datasets, data augmentation is also applied in the following steps:

- The input images are randomly resized within the scale of [0.5, 3] by preserving the aspect ratio. The images for the resizing operation are selected with a probability of 0.2.
- Next, the images are randomly rotated by an angle within $[-45°, 45°]$.
- Additionally, some images are sampled for random flip and crop within the scale of [0.5, 1]. The cropped image segments are resized to 224×224. The probability of 0.2 is used in the sampling step.
- Finally, the augmented examples having text instances smaller than half of the smallest text instances in the original dataset are filtered out.

Computing infrastructure: All simulations discussed in this paper were performed on NVIDIA Quadro P5000 GPU workstations with 32 GB RAM.

4.2. Baseline Evaluation

The baseline evaluation of the **MobileTDNet** architecture is performed without text attention blocks. The results are presented in Table 2. The traditional evaluation metrics of precision, recall, and F-score are used for bench marking. For the *MSRA-TD500* dataset, the evaluation protocol proposed in [69] is followed. The evaluation on *ICDAR* challenge datasets was performed in the manner specified in the challenge specification. Our sub-optimal implementation **MobileTDNet** took an average of 0.266 s to process the given input image (after required resizing). The analysis includes the computation time for the non-maximal suppression procedure.

Table 2. MobileTDNet baseline detection results.

Dataset	Precision	Recall	F-Score
ICDAR2013	0.930	0.845	0.885
ICDAR2015	0.882	0.826	0.853
MSRA-TD500	0.820	0.784	0.801
COCOText	0.622	0.592	0.606

Figure 3 shows the detection results for some difficult cases, which illustrate the detection challenges because of all variations in style, scale, and script. As observed, **MobileTDNet** efficiently captures text attributes in input images at all scales and in varying styles. It is remarkable that the **MobileTDNet** scene text detector, without an attention mechanism, performs comparable to many prominent scene text detectors (presented later in Table 3).

Figure 3. MobileTDNet output for sample difficult cases displayed in subfigures (**a**–**f**).

Table 3. MobileTDNet detection results with attention blocks: comparison with recent methods.

	ICDAR2013 Dataset			*ICDAR2015* Dataset			*MSRA-TD500* Dataset			*COCOText* Dataset		
Method	**Pre.**	**Rec.**	**F-Score**	**Pre.**	**Rec.**	**F-Score**	**Pre.**	**Rec.**	**F-Score**	**Pre.**	**Rec.**	**F-Score**
CTPN [40]	0.930	0.830	0.877	0.740	0.520	0.610	-	-	-	-	-	-
Text-block FCN [33]	0.880	0.780	0.830	0.710	0.430	0.540	0.830	0.670	0.740	-	-	-
RTN [39]	*0.940*	*0.890*	*0.910*	-	-	-	-	-	-	-	-	
Lyu et al. [51]	0.920	0.840	0.878	*0.895*	0.800	0.844	*0.880*	0.760	0.815	*0.620*	0.320	*0.425*
ATRR [43]	*0.937*	*0.897*	*0.917*	0.892	*0.860*	*0.876*	0.852	*0.821*	*0.836*	-	-	-
Mask TextSpotter [34]	**0.950**	*0.886*	*0.917*	*0.916*	0.810	0.860	-	-	-	-	-	-
Text-CNN [35]	0.930	0.730	0.820	-	-	-	0.760	0.610	0.690	-	-	-
TextBoxes++ [36]	0.910	0.840	0.880	0.878	0.785	0.829				*0.609*	*0.567*	*0.587*
PixelLink [46]	0.886	0.875	*0.881*	0.855	0.820	0.837	0.830	0.732	0.778	-	-	-
DB-ResNet-50 [72]	-	-	-	**0.918**	0.832	*0.873*	**0.915**	*0.792*	**0.849**	-	-	-
EAST [32]	-	-	-	0.833	0.783	0.807	*0.873*	0.674	0.761	0.504	*0.324*	0.395
Deng et al. [52]	-	-	-	-	-	-	-	-	-	*0.555*	**0.633**	*0.591*
SPCNET [47]	0.938	**0.905**	**0.921**	0.887	*0.858*	0.872	-	-	-	-	-	-
LOMO [53]	-	-	-	0.878	**0.876**	*0.877*	-	-	-	-	-	-
OPMP [73]	-	-	-	0.891	*0.855*	*0.873*	*0.860*	**0.834**	*0.847*	-	-	-
TextMountain [74]	-	-	-	0.885	0.842	0.863	-	-	-	-	-	-
MobileTDNet + Text Attention 1	0.941	0.853	0.895	0.903	0.842	0.871	0.843	0.802	0.822	0.627	0.596	0.611
MobileTDNet + Text Attention 1, 2 and 3	*0.947*	0.876	*0.910*	*0.913*	0.847	**0.879**	0.854	*0.807*	*0.830*	**0.631**	*0.603*	**0.617**

4.3. Effect of Skip Connections

The incorporation of skip connections is an important design component in the **MobileTDNet** architecture. The skip connections generate multiple levels of intermediate feature maps, i.e., $F^{\frac{1}{2}}$, $F^{\frac{1}{4}}$, $F^{\frac{1}{8}}$, $F^{\frac{1}{16}}$ and $F^{\frac{1}{32}}$, which are subsequently processed to generate predictions from the network. To analyze the contribution by skip connections, the ablation experiments are performed by gradually removing skip connections in the **MobileTDNet** architecture. The experiments focus on the *ICDAR2013* and *ICDAR2015* datasets where the former dataset is dominated by horizontal text appearances. Table 4 shows the summary of experiments also having the F-score difference with the baseline results given in Table 2.

Table 4. Analysis of skip connections on *ICDAR2013* and *ICDAR2015* datasets.

(a) *ICDAR201*					
Skip Connection: Feature Maps	**Pre.**	**Rec.**	**F-Score**	**F-Score Difference**	**Average Detection Time in Seconds**
$\{F^{\frac{1}{2}}, F^{\frac{1}{4}}, F^{\frac{1}{8}}, F^{\frac{1}{16}}\}$	0.835	0.814	0.823	*0.062*	0.278
$\{F^{\frac{1}{2}}, F^{\frac{1}{4}}, F^{\frac{1}{8}}\}$	0.787	0.763	0.775	0.110	0.265
$\{F^{\frac{1}{2}}, F^{\frac{1}{4}}\}$	0.735	0.704	0.719	0.166	0.246
$\{F^{\frac{1}{2}}\}$	0.705	0.678	0.691	0.194	0.231
$\{F^{\frac{1}{4}}, F^{\frac{1}{8}}, F^{\frac{1}{16}}, F^{\frac{1}{32}}\}$	0.842	0.801	0.821	*0.064*	0.267
$\{F^{\frac{1}{8}}, F^{\frac{1}{16}}, F^{\frac{1}{32}}\}$	0.777	0.751	0.764	0.121	0.280
(b) *ICDAR2015*					
Skip Connection: Feature Maps	**Pre.**	**Rec.**	**F-Score**	**F-Score Difference**	
$\{F^{\frac{1}{2}}, F^{\frac{1}{4}}, F^{\frac{1}{8}}, F^{\frac{1}{16}}\}$	0.825	0.788	0.806	*0.047*	
$\{F^{\frac{1}{2}}, F^{\frac{1}{4}}, F^{\frac{1}{8}}\}$	0.819	0.778	0.798	0.055	
$\{F^{\frac{1}{2}}, F^{\frac{1}{4}}\}$	0.684	0.627	0.654	0.199	
$\{F^{\frac{1}{2}}\}$	0.664	0.581	0.620	0.233	
$\{F^{\frac{1}{4}}, F^{\frac{1}{8}}, F^{\frac{1}{16}}, F^{\frac{1}{32}}\}$	0.829	0.796	0.812	*0.041*	
$\{F^{\frac{1}{8}}, F^{\frac{1}{16}}, F^{\frac{1}{32}}\}$	0.736	0.749	0.742	0.111	

As observed in the results, all skip connections contribute almost equally in the learning task since the text appearances in input images can be in many forms, shapes, and styles. Concurrently looking at F-score differences: for *ICDAR2015*, the skip connections with intermediate features captured at higher spatial resolutions are marginally less effective than connections to carry features captured at lower spatial resolutions. For the dataset, the absence of the $F^{\frac{1}{32}}$ feature map affects model learning more than the $F^{\frac{1}{2}}$ feature map. On the contrary, for *ICDAR2013*, the $F^{\frac{1}{2}}$ feature map is more important than that of $F^{\frac{1}{32}}$. The possible reason could be that the *ICDAR2015* consists of more variations including incidental text appearances with non-regular bounding boxes where skip connections capturing fine features are more important. In the next step, the visualization of detections for an example is investigated under two settings: (1) the network trained with $\{F^{\frac{1}{4}}, F^{\frac{1}{8}}, F^{\frac{1}{16}}, F^{\frac{1}{32}}\}$ and (2) with $\{F^{\frac{1}{8}}, F^{\frac{1}{16}}, F^{\frac{1}{32}}\}$ skip connections. As shown in Figure 4, the removal of the $\{F^{\frac{1}{4}}\}$ skip connection decreased the average IoU for the detections marginally. However, the careful observation of Figures 3f and 4a,b shows that detection quality significantly decreases with the removal of skip connections at higher spatial resolutions. This is also observed in the

overall results presented in Table 4. The average timing analysis for *ICDAR2013* shows that skip connections at lower spatial resolutions are relatively computationally intensive, as lower layers in the network contribute a bigger share of network weights.

(**a**) (**b**)

Figure 4. (**a**) shows the detections with $\{F^{\frac{1}{4}}, F^{\frac{1}{8}}, F^{\frac{1}{16}}, F^{\frac{1}{32}}\}$ skip connection, average IoU = 0.763; (**b**) shows the detections with $\{F^{\frac{1}{8}}, F^{\frac{1}{16}}, F^{\frac{1}{32}}\}$ skip connections, average IoU = 0.758.

4.4. Incorporation of Text Attention Blocks

Next, the contribution of text attention blocks in the **MobileTDNet** scene text detector is evaluated. The effect of text attention blocks into the detector network is analyzed in two steps. First, the **Text Attention 1** block is incorporated in the **MobileTDNet** network, which is followed by the incorporation of **Text Attention 2** and **Text Attention 3** blocks. Table 3 presents the summary of evaluation results which shows a consistent improvement in detection performance on test datasets with the incorporation of attention blocks in comparison with the baseline performance presented in Table 2. Table 3 also shows the performance of other state-of-the-art benchmarks. The best results for each measure are highlighted in bold. In addition, the next three best results are italicized. It is noteworthy that **MobileTDNet** achieves the best F-score on the *ICDAR2015* and *COCOText* datasets. Furthermore, our method has one of the top three F-scores for *ICDAR2013* and one for top three recall rates for the *MSRA-TD500* dataset.

Our analysis also focuses on the IoU measure on the *MSRA-TD500* dataset, which covers a wide range of variations in text appearances in a small collection of training and test sets. The overall performance of **MobileTDNet** on this dataset lags behind that of the *ICDAR2013* and *COCOText* datasets. To establish the contribution of attention blocks, the average IoU on the testing set is analyzed under different attention block settings, observing the average processing time to determine the average computational complexity. As shown in Table 5, the incorporation of attention blocks improved the average IoU measure on the testing set, although with a marginal difference. The average IoU improvement is reflected as an increase of 2.90% in the detection F-score in comparison with the baseline result presented in Table 2.

Table 5. Analysis of attention blocks on *MSRA-TD500* detections using IoU measure.

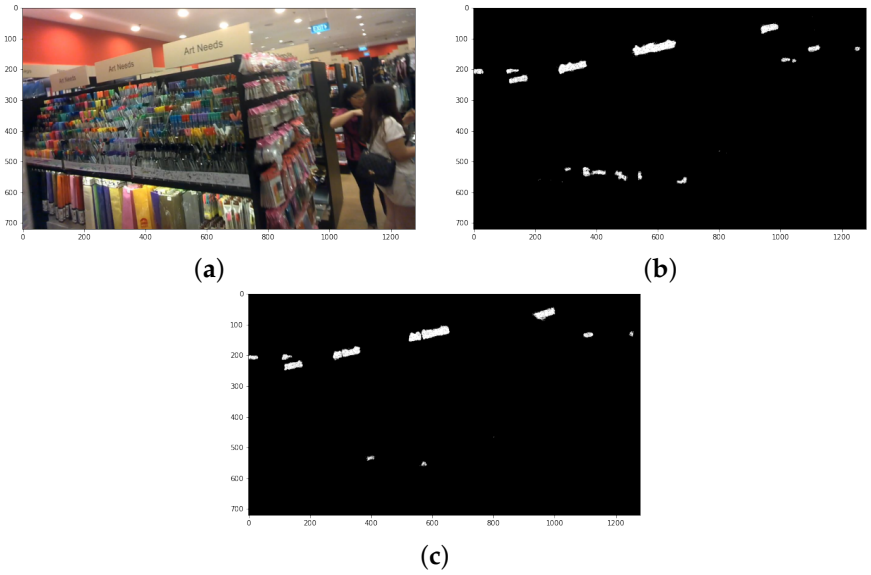
	W/O Attention	**Text Attention 1**	**Text Attention 1, 2 and 3**
Avg. IoU	0.772	0.786	0.793
Avg. processing time in seconds	0.385	0.415	0.486

Figures 5 and 6 further establish the efficacy of the proposed scene text detection pipeline. Figure 5 shows the quadrilateral bounding box detections for an example input, measuring the average IoU of detected quadrilateral boxes. In this case, the incorporation of **Text Attention 1, 2 and 3** blocks improved the average IoU by 2.20% with respect to the baseline, i.e., without text attention.

(**a**) (**b**) (**c**)

Figure 5. MobileTDNet text detection for an example: (**a**) shows detection without text attention blocks, average IoU = 0.859; (**b**) shows detection with **Text Attention 1** block, average IoU = 0.873; (**c**) shows detection with **Text Attention 1, 2 and 3** blocks, average IoU = 0.881.

Again, Figure 6 shows the filtered text confidence scope map for another example input with/without the incorporation of text attention blocks. The incorporation of **Text Attention 2 and 3** significantly improves the text/non-text segmentation by removing spurious text segments.



(**a**) (**b**)

(**c**)

Figure 6. Filtered text confidence score maps for the example in (**a**): the score map with **Text Attention 1** block shown in (**b**), and with **Text Attention 1, 2 and 3** blocks shown in (**c**).

5. Discussion

The **MobileTDNet** detection results in Table 3 present the best precision and F-score on the *COCOText* dataset with the next best recall rate. It is remarkable as the *COCOText* dataset is the largest of all datasets used in this work, where dataset images were collected for scene understanding research and consist of a large scale of diversity and variations in text appearances. Similarly, the proposed detector achieves the highest F-score for the *ICADR2015* dataset with the third best precision score. **MobileTDNet** improved on many prominent methods, including TextBoxes++ [36], EAST [32] and PixelLink [46], as shown in Table 3. Concurrently, the **MobileTDNet** performance does not surpass the state of the art on the *ICDAR2013* and *MSRA-TD500* datasets, although it is positioned in the top-3 in precision and F-score for *ICDAR2013* and in the recall rate for *MSRA-TD500*. The training set of these datasets is comparatively smaller, and it is likely that the **MobileTDNet** extracted features do not preserve sufficient representative attributes to learn an effective model. DB-ResNet-50 [72] performance closely matches with **MobileTDNet** on *ICDAR2015* {precision, recall, F-score difference: 0.005, −0.015, −0.006}; however, it performs better on the *MSRA-TD500* dataset {precision, recall, F-score difference: 0.061, −0.015, 0.019}. DB-ResNet-50 [72] and Mask TextSpotter [34] detect polygonal text segments in images and

require more complex processing in the convolutional network architecture. In contrast, the proposed method presents a simple convolutional architecture for scene text detection using depthwise separable convolutions, where the detector is designed for quadrilateral text box prediction in real-time applications.

Our method also achieves performance on par with SPCNET [47]. On *ICDAR2013*, **MobileTDNet** improves precision by 0.009 but trails on recall and F-score by 0.029 on recall and 0.011. Again, on *MSRA-TD500*, our method improves precision and F-score by 0.026 and 0.007 but trails in recall rate by 0.011. The SPCNET [47] architecture is designed for multi-scale feature extraction using a feature pyramid network (FPN) [75]. **MobileTDNet** also improves on precision and F-score measures on *ICDAR2015* in comparison with OPMP [73], which is also based on FPN-based multi-scale feature analysis. Our method also improves the recall rate on *MSRA-TD500* in comparison with OPMP by 0.027. The FPN is widely regarded as a state-of-the-art feature extractor based on deep convolutional networks with conventional convolution operation, with the goal to extract high-quality representative features. The proposed detector is designed for accurate scene text detection in images with all possible variations in text appearance, orientation, and background, and it has computationally efficient scene text detection suitable for real-time applications. With **MobileTDNet**, we achieve both objectives as a highly accurate scene text detector designed on depthwise separable convolution based deep convolutional network. Nevertheless, the direct prediction of quadrilateral bounding boxes is not accurate in the case of complex curved shapes in text appearances, which requires further investigation to split and merge bounding box predictions.

6. Conclusions and Future Work

The paper presented a novel convolutional net architecture for scene text detection, which can detect arbitrarily shaped text segments in scene images predicting quadrilateral bounding boxes with high accuracy. Our work demonstrated a novel application of a depthwise separable convolutions-based MobileNet in a U-shaped network design for a scene text detector with a text attention mechanism. The extensive experiments on standard public datasets establish the efficacy of the presented methodology in comparison with the state of the art where the results demonstrate that **MobileTDNet** outperformed many prominent scene text detection methods. **MobileTDNet** is designed for mobile and embedded applications as the target using the computationally efficient depthwise separable convolution operation. The incorporation of split and merge strategies into the proposed detector is the next problem to be explored in the proposed methodology. It can address the detection of complex shaped curved text instances, which is an important limitation of the present method. In addition, extending the proposed method into an end-to-end scene text recognition pipeline is another direction for future work.

Author Contributions: Investigation, L.V.L.; Project administration, E.H.; Supervision, E.H.; Writing—original draft, E.H. All authors have read and agreed to the published version of the manuscript.

Funding: The project was fully funded by the Kuwait Foundation for the Advancement of Sciences under the Project Grant PR19-18QI-01.

Data Availability Statement: The datasets used in the presented study are openly available at [65–67,69].

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Zhang, H.; Zhao, K.; Song, Y.Z.; Guo, J. Text extraction from natural scene image: A survey. *Neurocomputing* **2013**, *122*, 310–323. [CrossRef]
2. 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.

3. Neumann, L.; Matas, J. A Method for Text Localization and Recognition in Real-World Images. In Proceedings of the ACCV, Queenstown, New Zealand, 8–12 November 2010.
4. Lienhart, R.; Stuber, F. Automatic text recognition in digital videos. In Proceedings of the Electronic Imaging, San Jose, CA, USA, 28 January–2 February 1996.
5. Cai, M.; Song, J.; Lyu, M.R. A new approach for video text detection. In Proceedings of the International Conference on Image Processing, Rochester, NY, USA, 22–25 September 2002; Volume 1. [CrossRef]
6. Agnihotri, L.; Dimitrova, N. Text detection for video analysis. In Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL'99), Fort Collins, CO, USA, 22 June 1999; pp. 109–113. [CrossRef]
7. Ezaki, N.; Bulacu, M.; Schomaker, L. Text detection from natural scene images: Towards a system for visually impaired persons. In Proceedings of the of the 17th International Conference on Pattern Recognition (ICPR 2004), Cambridge, UK, 26 August 2004; Volume 2, pp. 683–686. [CrossRef]
8. Neumann, L.; Matas, J.E.S. Real-time scene text localization and recognition. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3538–3545.
9. Yin, X.; Yin, X.; Huang, K. Robust Text Detection in Natural Scene Images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 970–983. [PubMed]
10. Cho, H.; Sung, M.; Jun, B. Canny Text Detector: Fast and Robust Scene Text Localization Algorithm. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3566–3573. [CrossRef]
11. Li, Y.; Lu, H. Scene text detection via stroke width. In Proceedings of the of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 11–15 November 2012; pp. 681–684.
12. Gomez, L.; Karatzas, D. MSER-based real-time text detection and tracking. In Proceedings of the 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 3110–3115.
13. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
14. Liu, X.; Meng, G.; Pan, C. Scene text detection and recognition with advances in deep learning: A survey. *Int. J. Doc. Anal. Recognit.* **2019**, *22*, 143–162. [CrossRef]
15. Long, S.; He, X.; Yao, C. Scene text detection and recognition: The deep learning era. *Int. J. Comput. Vis.* **2021**, *129*, 161–184. [CrossRef]
16. Available online: https://rrc.cvc.uab.es/ (accessed on 8 May 2022).
17. Available online: https://paperswithcode.com/sota/scene-text-detection-on-msra-td500 (accessed on 8 May 2022).
18. Available online: https://paperswithcode.com/sota/scene-text-detection-on-coco-text (accessed on 8 May 2022).
19. Howard, A.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
20. Ba, J.; Mnih, V.; Kavukcuoglu, K. Multiple object recognition with visual attention. *arXiv* **2014**, arXiv:1412.7755.
21. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer Networks. In Proceedings of the of the 28th International Conference on Neural Information Processing Systems (NIPS'15), Montreal, QC, Canada, 7–12 December 2015; Volume 2, pp. 2017–2025.
22. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
23. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. Bam: Bottleneck attention module. *arXiv* **2018**, arXiv:1807.06514.
24. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. *arXiv* **2019**, arXiv:1910.03151.
25. Delakis, M.; Garcia, C. Text Detection with Convolutional Neural Networks. In Proceedings of the VISAPP (2), Madeira, Portugal, 22–25 January 2008; pp. 290–294.
26. Wang, T.; Wu, D.J.; Coates, A.; Ng, A.Y. End-to-end text recognition with convolutional neural networks. In Proceedings of the of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 11–15 November 2012; pp. 3304–3308.
27. Jaderberg, M.; Vedaldi, A.; Zisserman, A. Deep Features for Text Spotting. In Proceedings of the ECCV, Zurich, Switzerland, 6–12 September 2014.
28. Girshick, R.B. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
29. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object Detection via Region-based Fully Convolutional Networks. In Proceedings of the Advances in Neural Information Processing Systems 29 (NIPS 2016), Barcelona, Spain, 5–10 December 2016; pp. 379–387.
30. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.; Berg, A.C. SSD: Single Shot MultiBox Detector. *arXiv* **2015**, arXiv:1512.02325.
31. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
32. Zhou, X.; Yao, C.; Wen, H.; Wang, Y.; Zhou, S.; He, W.; Liang, J. EAST: An Efficient and Accurate Scene Text Detector. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2642–2651. [CrossRef]

33. Zhang, Z.; Zhang, C.; Shen, W.; Yao, C.; Liu, W.; Bai, X. Multi-oriented Text Detection with Fully Convolutional Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4159–4167. [CrossRef]
34. 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 of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
35. 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] [PubMed]
36. Liao, M.; Shi, B.; Bai, X. TextBoxes++: A Single-Shot Oriented Scene Text Detector. *IEEE Trans. Image Process.* **2018**, *27*, 3676–3690. [CrossRef] [PubMed]
37. Qin, X.; Zhou, Y.; Guo, Y.; Wu, D.; Tian, Z.; Jiang, N.; Wang, H.; Wang, W. Mask is All You Need: Rethinking Mask R-CNN for Dense and Arbitrary-Shaped Scene Text Detection. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 414–423.
38. Liao, M.; Shi, B.; Bai, X.; Wang, X.; Liu, W. TextBoxes: A Fast Text eastDetector with a Single Deep Neural Network. *arXiv* **2016**, arXiv:1611.06779.
39. 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.
40. Tian, Z.; Huang, W.; He, T.; He, P.; Qiao, Y. Detecting Text in Natural Image with Connectionist Text Proposal Network. *arXiv* **2016**, arXiv:1609.03605.
41. 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.
42. Yao, C.; Bai, X.; Shi, B.; Liu, W. Strokelets: A learned multi-scale representation for scene text recognition. In Proceedings of the of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 4042–4049.
43. Wang, X.; Jiang, Y.; Luo, Z.; Liu, C.; Choi, H.; Kim, S. Arbitrary Shape Scene Text Detection With Adaptive Text Region Representation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 6442–6451.
44. Xing, L.; Tian, Z.; Huang, W.; Scott, M.R. Convolutional character networks. In Proceedings of the of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9126–9136.
45. Zhong, Y.; Cheng, X.; Chen, T.; Zhang, J.; Zhou, Z.; Huang, G. PRPN: Progressive region prediction network for natural scene text detection. *Knowl. Based Syst.* **2022**, *236*, 107767. [CrossRef]
46. Deng, D.; Liu, H.; Li, X.; Cai, D. Pixellink: Detecting scene text via instance segmentation. In Proceedings of the of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
47. Xie, E.; Zang, Y.; Shao, S.; Yu, G.; Yao, C.; Li, G. Scene text detection with supervised pyramid context network. In Proceedings of the of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 9038–9045.
48. 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, HI, USA, 7–11 January 2019; pp. 764–772.
49. He, W.; Zhang, X.Y.; Yin, F.; Liu, C.L. Deep Direct Regression for Multi-Oriented Scene Text Detection. *arXiv* **2017**, arXiv:1703.08289.
50. Shi, B.; Bai, X.; Belongie, S. Detecting Oriented Text in Natural Images by Linking Segments. In Proceedings of the of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
51. 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 of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7553–7563.
52. 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]
53. 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. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 10544–10553. [CrossRef]
54. Hu, H.; Zhang, C.; Luo, Y.; Wang, Y.; Han, J.; Ding, E. WordSup: Exploiting Word Annotations for Character Based Text Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4950–4959.
55. Fragoso, V.; Gauglitz, S.; Zamora, S.; Kleban, J.; Turk, M. TranslatAR: A mobile augmented reality translator. In Proceedings of the 2011 IEEE Workshop on Applications of Computer Vision (WACV), Kona, HI, USA, 5–7 January 2011; pp. 497–502.
56. Petter, M.; Fragoso, V.; Turk, M.; Baur, C. Automatic text detection for mobile augmented reality translation. In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 48–55.
57. Yi, C.; Tian, Y. Scene text recognition in mobile applications by character descriptor and structure configuration. *IEEE Trans. Image Process.* **2014**, *23*, 2972–2982. [CrossRef]

58. Shivakumara, P.; Wu, L.; Lu, T.; Tan, C.L.; Blumenstein, M.; Anami, B.S. Fractals based multi-oriented text detection system for recognition in mobile video images. *Pattern Recognit.* **2017**, *68*, 158–174. [CrossRef]
59. Fu, K.; Sun, L.; Kang, X.; Ren, F. Text Detection for Natural Scene based on MobileNet V2 and U-Net. In Proceedings of the 2019 IEEE International Conference on Mechatronics and Automation (ICMA), Tianjin, China, 4–7 August 2019; pp. 1560–1564.
60. Shrivastava, A.; Gupta, A.; Girshick, R. Training region-based object detectors with online hard example mining. In Proceedings of the of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 761–769.
61. Lin, T.Y.; Goyal, P.; Girshick, R.B.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]
62. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
63. Neubeck, A.; Van Gool, L. Efficient non-maximum suppression. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; Volume 3, pp. 850–855.
64. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS–improving object detection with one line of code. In Proceedings of the of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5561–5569.
65. Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; i Bigorda, L.G.; 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.
66. 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), Tunis, Tunisia, 23–26 August 2015; pp. 1156–1160.
67. 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.
68. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common Objects in Context. *arXiv* **2014**, arXiv:1405.0312.
69. Yao, C.; Bai, X.; Liu, W.; Ma, Y.; Tu, Z. Detecting texts of arbitrary orientations in natural images. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1083–1090. [CrossRef]
70. Tieleman, T.; Hinton, G. Lecture 6. In *COURSERA: Neural Networks for Machine Learning*; Coursera: Mountain View, CA, USA, 2012.
71. Bengio, Y. Practical recommendations for gradient-based training of deep architectures. In *Neural Networks: Tricks of the Trade*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 437–478.
72. Liao, M.; Wan, Z.; Yao, C.; Chen, K.; Bai, X. Real-time Scene Text Detection with Differentiable Binarization. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
73. Zhang, S.; Liu, Y.; Jin, L.; Wei, Z.; Shen, C. OPMP: An Omnidirectional Pyramid Mask Proposal Network for Arbitrary-Shape Scene Text Detection. *IEEE Trans. Multimed.* **2021**, *23*, 454–467. [CrossRef]
74. Zhu, Y.; Du, J. TextMountain: Accurate scene text detection via instance segmentation. *Pattern Recognit.* **2021**, *110*, 107336. [CrossRef]
75. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]