Robust Scene Text Detection for Partially Annotated Training Data

Prateek Keserwani, Rajkumar Saini, Marcus Liwicki and Partha Pratim Roy

***Abstract*—This article analyzed the impact of training data containing un-annotated text instances, i.e., partial annotation in scene text detection, and proposed a text region refinement approach to address it. Scene text detection is a problem that has attracted the attention of the research community for decades. Impressive results have been obtained for fully supervised scene text detection with recent deep learning approaches. These approaches, however, need a vast amount of completely labeled datasets, and the creation of such datasets is a challenging and time-consuming task. Research literature lacks the analysis of the partial annotation of training data for scene text detection. We have found that the performance of the generic scene text detection method drops significantly due to the partial annotation of training data. We have proposed a text region refinement method that provides robustness against the partially annotated training data in scene text detection. The proposed method works as a two-tier scheme. Text-probable regions are obtained in the first tier by applying hybrid loss that generates pseudo-labels to refine text regions in the second-tier during training. Extensive experiments have been conducted on a dataset generated from ICDAR 2015 by dropping the annotations with various drop rates and on a publicly available SVT dataset. The proposed method exhibits a significant improvement over the baseline and existing approaches for the partially annotated training data.**

***Index Terms*—Partial annotation, Scene text detection, Text region refinement, Pseudo-labeling**

I. Introduction

The textual and written content is one of the prominent reasons behind the organized and civilized human society. The textual information's impact is spread across the literature on the utility of text to visualize the brand, location information, number plate of vehicles, and house number. This written content spread in our surroundings is termed as scene text [1]. Scene text detection is an important problem and has become an essential aspect for scene understanding [2], [3]. Impressive results have been obtained on the scene text by fully supervised text detection (FSTD) with the help of deep learning methods [4]. The scene text always appears in an image in a group and most probably present in specific places [5]. This grouping of text in an image makes the annotation a time-consuming and expensive task. The fully annotated training data drives the performance of the fully supervised deep learning methods [6], [7]. However, if some text instances are not annotated in a training image, this scenario comes under the umbrella of the partial annotation issue. Partial annotation introduces unnecessary noise in training data. The impact of the partially annotated training data on the scene text is still unaddressed and needs to be analyzed.

Fig. 1: Samples of partially annotated images in publicly available SVT dataset. The provided annotations are shown by yellow color, whereas the missing label is shown by red color bounding boxes.

The two main factors behind deep learning success in various domains are: (1) deeper neural network architecture design and (2) training on massive datasets. In massive datasets, accurate and quality annotation generation is a tedious and time-consuming task. As the dataset size scales up, the fully manual annotation by a single person is not an efficient solution. In such a case, the involvement of crowd-sourced humans for the labeling is required. However, such an annotation method is time-consuming and costly. In the word spotting problem, dataset such as SVT [8] has tagged only some of the text instances. In Fig. 1, some samples of the SVT dataset with partial ground truth annotation are shown. It is challenging to train a text detector on this partially annotated dataset. The method which can get trained efficiently on this partially annotated dataset helps to reduce the manual effort and save a huge amount of time for complete annotation.

Recently, the impact of quality annotation on deep learning has become a new research direction. There are many studies on the impact of noisy labels on deep learning models for various problems [9]–[12]. In face recognition, label noise significantly drops the model's performance. The training on a small amount of clean annotation produces better results than a larger noisy dataset [9]. In [10], the training is based on noisy data, the model overfits the noisy data, and the model highly depends on label noise. They empirically found that the deep learning models are more sensitive to concentrated label noise than the spread label noise across the training data. In [12], the impact of noisy class label on semantic segmentation has been discussed and showed the impact of noise-robust loss's role for better learning. In [11], the imprecise bounding

Prateek Keserwani and Partha Pratim Roy are with Indian Institute of Technology, Roorkee, India (email id: pkeserwani@cs.iitr.ac.in, partha@cs.iitr.ac.in)

Rajkumar Saini and Marcus Liwicki are with EISLAB Machine Learning Luleå , University of Technology Luleå, Sweden (email id: rajkumar.saini@ltu.se, marcus.liwicki@ltu.se)

Manuscript received .....; revised .....

box along with the incorrect class label imposes a bigger challenge for object detection. The negative impact of noisy labels on deep models for real-world applications has also been reported in [13] and [14]. In the context of degradation in annotation quality, the partial annotations introduced noise. All the existing text detection approaches highly rely on the hypothesis that the text annotations are correct and complete.

In prior research, it has been observed that the partially annotated training data hampers the performance of the fully supervised object detection [15]–[17]. The impact of missing labels is due to applying the loss function to the object of interest. From the studies, it has been found that the log-loss and exponential loss are not suitable for noisy annotations [18], [19]. The basic approaches for object detection under the constraint of partially annotated training data include re-weighting the hard negatives and ignoring the wrong gradient of false-negative regions to reduce its impact. In our considered assumption for scene text detection, the provided text instances are perfect, and the noise occurs by assigning some text location as a background. The ambiguity of the text instance as a background leads to the difficulty of learning the text features. This visual similarity of the text strokes with the background makes text detection different from object detection under partial annotation.

To handle the ambiguity between unlabelled text and background pixels, we have proposed a text region refinement module. The proposed method works on two levels. It first performs the prediction of text-probable regions by using a hybrid loss to obtain the text-biased segmentation and avoids the log-loss due to its adverse impact in case of noisy labels [18], [19]. The pseudo-labels are generated using the provided ground truth and the text-biased segmentation. For the second level, the text region refinement is performed using pseudo-labels. In this work, we propose a method for text detection in the presence of a partial annotation for training and analyze the impact of the partial annotation on existing deep learning-based approaches of FSTD.

The major contribution of this work is four-fold:

- To the best of our knowledge, this is the first work that addresses the impact of the noise due to partially annotated training data on the scene text detection problem.
- The presented work analyzes the robustness of the scene text detector under the constraint of partially annotated training data.
- Proposed a text region refinement method for the robust scene text detection.
- Extensive analysis of the proposed method on the generated and publicly available partially annotated dataset.

The rest of the paper is organized into four sections. Section II is concerned with the work related to the posed problem. The baseline and the proposed method are discussed in Section III. The experiment and the analysis of the proposed method are presented in Section IV. Finally, the conclusion and future work are mentioned in Section V.

II. Related Work

Existing approaches of text detection under the assumption of fully annotated training samples have achieved plausible performance on many public datasets [20]–[27]. In [28], the text detection-based text concealment has been performed with the help of fully labeled and clean synthetic data with no real-world data. Some of the existing methods [29] on scene text detection has used the weakly supervised learning [30]. In [29], the fully annotated dataset along with the weakly annotated images has taken for scene text reading. The weak annotation consists of text transcription without the bounding box information. In [29], it is well defined that either the image is completely labeled or weakly labeled i.e. annotation is noise free. But in the case of the partial annotated data the ground truth is noisy which makes the learning difficult. The partial annotation is also different from the semi-supervised setting [31], [32]. In the case of a semi-supervised setting, a set of images is completely annotated, and there is another set of unlabeled images. Semi-supervised learning aims to harness the additional unlabeled data to improve performance. However, in our case, the bounding box is missing from random images, and hence, there is no guarantee of complete annotation.

To the best of our knowledge, no work wisely considered the partial annotation problem and handled the presence of noise in training data due to partially annotated text instances in scene text detection. Hence, the impact of partially annotated training data is not well explored in scene text detection. On the other hand, a few works exist for object detection under the partially annotated training data constraint [15]–[17], [33]–[36]. In [15], an overlap-based soft sampling technique has been proposed for robust object detection and is included as a part of training on Faster-RCNN [37]. They reduced the gradient based on the intersection with positive samples. They decayed the gradient of each ROI according to its overlap with the positive samples. For this decay, they have used the Gompertz function. Low weight is assigned to the background pixels and high weight to the positive and hard negative ROIs. Hence, in the case of ambiguous regions, the gradient flow is well regulated. In [16], [33], the authors have proposed combining the fully-supervised and weakly-supervised learning to solve the missing annotation problem. [16] is based on teacher-student learning. The teacher detector provides the pseudo label to the student detector for training. In [33], the authors have proposed a pseudo ground truth mining procedure for missing annotations, then combined the available and mined ground truth for fully supervised training. For further improvement, incremental training has been added to the proposed approach. In [17], authors have introduced pseudo-label guided sampling to handle missing annotation issues. The same network is trained twice. The prediction from the first has been used as a pseudo label for the second time learning. In [34], the re-calibration loss has been introduced inspired by the focal loss, which automatically re-calibrates the loss signal as per the IoU threshold. [35] considered the missing annotation problem as a positive un-label problem. In [36], a concept of co-mining has been introduced. It consists of a Siamese network that predicts the pseudo-label for each other. The original and augmented images have been used as input to the Siamese network.

The above-mentioned object detection methods for missing

annotation due to partially annotated training data are either re-weighting the loss for hard negatives or ignoring the regions to avoid the false gradient due to false negatives. The proposed method has combined both approaches by reducing the impact of the wrong gradient with the use of hybrid loss and then generating the pseudo-label for the ambiguous regions to avoid the wrong gradient flow for false negative. In the past, for visual understanding, extra cues have been used by using the multiple knowledge representation [38]. In the proposed work, a pseudo-label has been generated to guess the textual region that is part of the background as per the provided ground truth. It helps to avoid incorrect gradients and helps to achieve robustness. Hence, the pseudo-label generation can be considered as an extra cue for the refinement block to avoid certain portions of images. The proposed pseudo-label uses the cue from the text-biased segmentation to the refined segmentation for robust representation. This idea of leveraging the cues for learning is similar to [38] but on deep representation obtained from the shared backbone, i.e., single knowledge representation.

III. Methods

This section discusses the baseline approach for text detection under the partially annotated training data hypothesis in detail. Afterward, an improved version known as the text region refinement method has been explained, whose design is based on the baseline approach.

A. Baseline Method

For the development of the baseline method, we have utilized the text descriptor known as maximally stable extremal region (MSER) [39]. The MSER has been proved a good text descriptor [5], [40]–[42]. The advantage of using the MSER is its capability to identify challenging text patterns. The MSER is an over-segmentation algorithm to extract connected components representing regions with a similar property as a textual region. It helps to get a good recall for text detection. In partial annotation, many textual regions are not labeled; hence, these regions are considered as the background regions by existing text detection methods. This missing label disrupts the convolutional neural network training for text detection. Anyhow, the few background regions that possess the text's property should be dropped during training. It helps to ignore the un-annotated false negatives in the training phase. It provides a cleaner gradient on the network. The background regions, which are a subset of the MSERs, are considered a good prior for considering the ambiguous region from a background that can be a text region. The effectiveness of considering the background intersected MSERs as the ambiguous regions are illustrated in Fig. 2. We have chosen the EAST [26] text detector for this baseline approach, which considers the background overlapping MSERs as don't care regions during training. It is considered a baseline method for text detection under the hypothesis of partial annotation. The advantage of this method is that it provides a cleaner gradient to the network, whereas at the same time, it ignores significant regions of images during training.

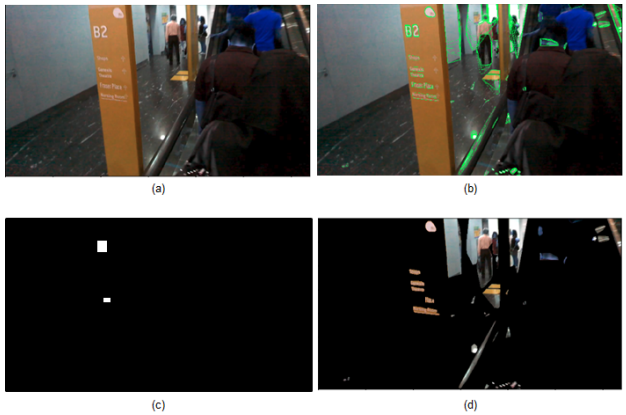


Fig. 2: (a) Image, (b) MSERs highlighted by green-colored enclosed regions, (c) Ground truth with missing annotations (d) Don't care regions while training.

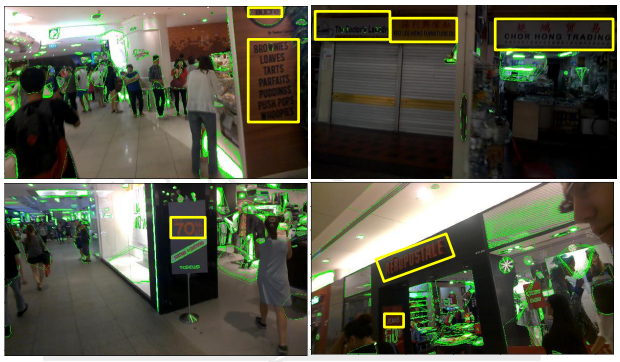


Fig. 3: Sample of images in which MSER is not able to detect the text regions. The MSERs are shown by green-colored enclosed regions, whereas the undetected text regions are shown by yellow-colored bounding boxes.

B. Text region refinement method

This section describes the details of the proposed method for text detection, which is an effective solution for the partially annotated training data. The proposed method is based on the baseline method. The baseline method is highly dependent on three factors. The first factor is the MSER's performance; the second is that the MSER itself decides the upper bound of the baseline approach, and the third is to ignore many un-annotated true positives while sustaining true negatives for learning. Some of the cases in which MSER cannot detect text and consider it as a background are shown in Fig. 3. We have proposed a text region refinement method to counter this high dependence of the baseline on the low-level MSER. It also includes more true negatives compared to the baseline approach. A two-level segmentation refinement module has been proposed to handle the missing annotation and avoid the dependence of the MSERs. The details of the method have been covered in the following subsections.

1) Network Design: The proposed method is based on the EAST text detector. VGGNet [43] has been used as a backbone. First, the architecture down-samples the features and learns the high-level representations, then progressively

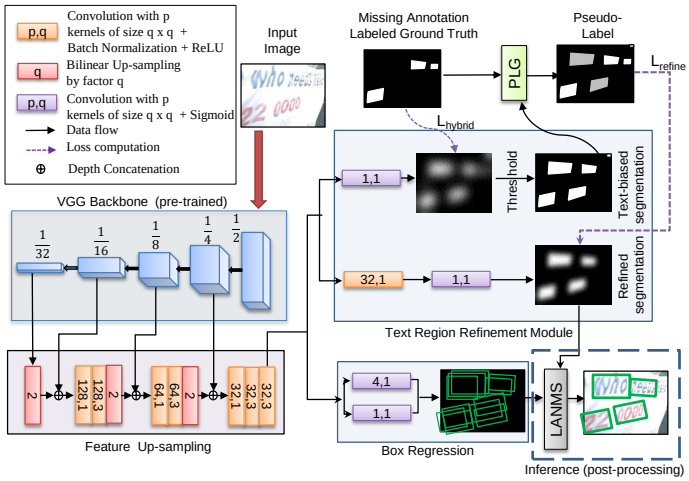


Fig. 4: The proposed method consists of the VGGNet backbone with a feature up-sampling branch with a box regression and text region refinement module. The text region refinement module first predicts the text-biased segmentation map from which the pseudo-labels have been generated using PLG (Pseudo-Label Generation) block for refined segmentation. Finally, the post-processing of LANMS (locality-aware non-maximal suppression) is applied on the bounding boxes using refined segmentation as a score of the boxes.

up-samples the features. The output consists of two heads. The first is for the bounding box regression, and the second is for the segmentation. The regression head is the same as in EAST. The segmentation head has been modified to give robustness against the wrong labeling of text instances as a background for handling the missing annotation. The detailed diagram of the proposed method is shown in Fig. 4. In the EAST text detector, the box regression learns separately, and the confidence of each box is taken from the segmentation. Our method only applies regression loss on the bounding box that appears on the available ground truth locations. The remaining bounding boxes are considered don't care cases. Hence, the training of the regression head will not get hampered due to missing annotations. Although the wrong ground truth for text pixels as a background disrupts the training, segmentation is the main issue behind this drop in performance. It motivated us to introduce the text region refinement module. The text region refinement module works in two-tier segmentation, namely, text-biased segmentation and refined segmentation. For the text-biased segmentation, the hybrid loss has been applied. A pseudo-label has been generated from this text-biased segmentation for the background pixels. For refined segmentation, the pseudo-labeled ground truth background pixels are sampled and considered as background ground truth.

For the segmentation in the standard EAST text detection, balanced cross-entropy has been used. However, from prior researches, it has been found that the log-based losses are not suitable for noisy data-based training [44]. Hence, in this work, we do not rely on the balanced cross-entropy loss function. In place of balanced cross-entropy, we have used the dice loss. Dice loss is independent of the log-loss. The provided annotated text as ground truth is correct, but some of the text instances are missing. Therefore, some of the text pixels are assigned as a background in the training phase. Using the dice loss would have given an equal penalty for both classes, and due to missing annotation, some text is considered background, which leads to the flow of the wrong gradient in the network. Inspired from the DB loss [45], [46], which combines the dice loss with boundary loss for a highly unbalanced dataset, we also combined the dice loss with some other loss to tackle noise. For the case of DB loss, the boundary loss is applied for the object contour boundary adjustment, and the dice loss is applied for the foreground and background. Hence, in the case of background pixels with noise, the boundary loss may impact the ground truth text regions, which is already correct but ineffective for background pixels. Dice loss handles the background regions and is not suitable for noisy background pixels. Hence, separate losses can be used for text regions and background regions. Considering the background regions, the noise tolerance loss function is good to avoid the wrong gradient in the neural network. From the findings of [44], the L1 loss is more noise-tolerant than the log-loss and L2 loss. Hence, the dice loss has been used for the foreground (i.e., text pixel), whereas L1 loss has been employed for the background pixels. The dice loss gives a considerable value for a small number of miss-classified pixels. In scene text, only a few pixels are ambiguous, even for high drop rates. Therefore, the wrong gradient is only concerned with relatively fewer pixels than correct background pixels. The noise tolerance property of L1 loss helps to counter wrong gradients up to a certain extent. Combining the dice and L1 loss helps avoid the wrong gradients and uplift the recall of text detection. The regions that can be text in the background are more likely to achieve higher activation, producing the text-biased map for text segmentation.

The text-biased segmentation helps to generate the pseudo-label for the background pixels. The provided ground truth for background pixels is then updated with the help of text-biased segmentation. The regions of the background in the provided annotation, which are found to be the text regions in the text-biased segmentation, are assigned as a new category known as ambiguous region. These ambiguous regions are used as don't care regions for the refined segmentation. This pseudo-labeling helps to neglect the ambiguous regions during training to avoid the wrong gradient and train robustly against the noisy annotation. Finally, the standard dice loss has been used at the refined segmentation.

2) Text region refinement module: The text region refinement module has two segmentation branches. The first branch estimates the text-biased segmentation, which is further refined by the refinement branch. The text-biased segmentation and the refined segmentation maps are at the one-fourth scale of the input image. These segmentation branches have used the sigmoid activation map to estimate the value in the range of $[0, 1]$. For the segmentation ground truth, we have considered the shrunk text bounding box as in [26] and [20] methods.

Text-biased segmentation The filled polygon map of the

shrunk polygon is considered as ground truth ($\mathcal{G}_s$) for the text-biased segmentation. The estimated text-biased segmentation $\mathcal{S}_c$ consists of regions that have potential for text. The following equation obtains the possible regions:

$$S_r = \begin{cases} 1, & \text{if } S_c > T \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where T is the threshold value in range $[0, 1]$. The threshold T helps to extract the text-probable regions.

Refined segmentation: The refined segmentation has been done between the text and non-text regions while considering the ambiguous regions as don't care regions to avoid wrong gradients during training. The pseudo-label generation (PLG) block helps to identify the text-probable regions (i.e. ambiguous region) from the background ground truth. The pseudo-label is generated with the thresholded text-biased segmentation (S_r), text-biased segmentation ground truth ($\mathcal{G}_s$), and segmentation map of original bounding boxes ($\mathcal{G}_o$) as the following equations:

$$\mathcal{G}_p = \begin{cases} -1, & \text{if } (1 - \mathcal{G}_s) \times (1 - \mathcal{G}_b) \times S_r = 1 \\ S_r, & \text{otherwise} \end{cases} \tag{2}$$

$$\mathcal{G}_b = \mathcal{G}_o - \mathcal{G}_s \tag{3}$$

The category -1 is used as a category that is supposed to be ambiguous and considered as don't care regions for refined segmentation.

3) Box Regression: For handling the multi-oriented bounding boxes, a preferred box representation is a rotated rectangle, which is used in this work. This rotated bounding box representation predicts the four perpendicular distances from each spatial location inside the bounding box to the rectangle's four sides and an angle. These five regression values can be considered as an ordered pair (d_1, d_2, d_3, d_4, a) for rotated rectangle representation. The following equation gives the activation function for regression:

$$l_i = sigmoid(d_i) * \gamma \tag{4}$$

$$\theta = \frac{1}{2}(2 \times sigmoid(a) - 1) * \pi \tag{5}$$

where γ and π are the scaling factors. The γ scales up the prediction to the range of the cropping window used during training. The π is the scaling factor to convert the angle into radian.

4) Loss Function: The network consists of two tasks, namely, regression and segmentation. The regression includes the rotated bounding box estimation. The rotated bounding box loss is considered as a horizontal bounding box loss ($\mathcal{L}_{box}$) along with the angle loss ($\mathcal{L}_{angle}$). For the horizontal bounding box loss between the predicted bounding box B and the ground truth bounding box $\hat{B}$, the intersection over union loss has been used, and the equations for the computation of the $\mathcal{L}_{box}$ are as follows:

$$\mathcal{L}_{box} = -log\left(\frac{\mathcal{A}(B \cap \hat{B}) + 1}{\mathcal{A}(B \cup \hat{B}) + 1}\right) \tag{6}$$

$$\mathcal{A}(B \cap \hat{B}) = (min(d_3, \hat{d_3}) + min(d_4, \hat{d_4})) * (min(d_1, \hat{d_1}) + min(d_2, \hat{d_2})) \tag{7}$$

$$\mathcal{A}(B \cup \hat{B}) = ((\hat{d_1} + \hat{d_2}) * (\hat{d_3} + \hat{d_4})) + ((d_1 + d_2) * (d_3 + d_4)) - A(B \cap \hat{B}) \tag{8}$$

The loss function for the computation of the angle loss between the predicted angle (θ) and the ground truth angle ($\hat{\theta}$) is given by the following equation:

$$\mathcal{L}_{angle} = 1 - cos(\theta - \hat{\theta}) \tag{9}$$

The total regression loss is the weighted sum of the horizontal bounding box loss ($\mathcal{L}_{box}$) and angle loss ($\mathcal{L}_{angle}$) given by the equation:

$$\mathcal{L}_{reg} = \lambda_1 \mathcal{L}_{box} + \lambda_2 \mathcal{L}_{angle} \tag{10}$$

where λ_1 and λ_2 used in this work are 1 and 10, respectively. The segmentation branch has two losses. First loss is for text-biased segmentation. This text-biased segmentation is achieved by the hybrid loss defined as:

$$\mathcal{L}_{hybrid} = 1 - \frac{2\sum \mathbb{1}^{\mathcal{G}_s}(\mathcal{G}_s \mathcal{S}_c)}{\sum \mathbb{1}^{\mathcal{G}_s}\mathcal{G}_s + \sum \mathbb{1}^{\mathcal{G}_s} s\mathcal{S}_c} + \frac{1}{\sum |1 - \mathcal{G}_s|} \sum \mathbb{1}^{1-\mathcal{G}_s} |\mathcal{G}_s - \mathcal{S}_c| \tag{11}$$

The second loss is for the refined segmentation, which has been used as final segmentation and is given by the following equation:

$$\mathcal{L}_{refine} = 1 - \frac{2\sum \mathbb{1}^{\mathcal{G}_p \neq -1}(\mathcal{G}_p \mathcal{S}_p)}{\sum \mathbb{1}^{\mathcal{G}_p \neq -1}\mathcal{G}_p + \sum \mathbb{1}^{\mathcal{G}_p \neq -1}\mathcal{S}_p} \tag{12}$$

The total segmentation loss ($\mathcal{L}_{seg}$) is the weighted sum of the $\mathcal{L}_{hybrid}$ and $\mathcal{L}_{refine}$ given by the following equation:

$$\mathcal{L}_{seg} = \lambda_3 \mathcal{L}_{hybrid} + \lambda_4 \mathcal{L}_{refine} \tag{13}$$

where the value of λ_3 and λ_4 are fixed to the value 1 in this work.

5) Post-processing: The bounding box generated by the proposed method is highly correlated with nearby pixels. These correlated bounding boxes overlap with each other. These overlapping bounding boxes can be merged using the standard non-maximal suppression (NMS) algorithm. However, it has been found in the [26] that the locality-aware non-maximal suppression (LANMS) variant of NMS produces better results in this configuration. Hence, locality-aware non-maximal suppression has been used in this method as a post-processing method to suppress the overlapping bounding boxes. The segmentation produced after the refined segmentation has been used as a scoring function for the bounding boxes. If two bounding boxes overlap under some threshold, these two bounding boxes are combined to generate a new bounding box.

Algorithm 1 Creation of partial label training dataset.

procedure DATASET(D, d_r)
  $counter \leftarrow 0$
  while $counter \leq |D|$ **do**
    $Annotation(D[counter], d_r)$
    $counter \leftarrow counter + 1$
procedure ANNOTATION(A, d_r)
  $counter \leftarrow 0$
  $new\_ann \leftarrow [\,]$
  while $counter \leq |A|$ **do**
    $rand \leftarrow Random()$ ▷ Uniform distribution
    if $rand \leq d_r$ **then**
      $new\_ann.append(A[counter])$
    $counter \leftarrow counter + 1$
  $Write\_Annoation(new\_ann))$ ▷ Write in a file

IV. Experiments and Results

A. Partially Annotated Dataset

Generated: For the analysis of the above-stated problem, a dataset has been created by randomly dropping the boxes from the fully annotated public dataset. The algorithm to create the missing label dataset with a drop rate of d_r from any fully annotated dataset is given by Algorithm 1. For generating a partially annotated dataset, ICDAR 2015 [47] has been chosen. The number of training and the testing images are 1000 and 500, respectively. The number of text instances in training images is 11886. After dropping the 20%, 40%, and 60% bounding boxes, the number of ambiguous pixels on the training images is 0.50%, 1.04%, and 1.50%, respectively.

Publicly available: The SVT [8] dataset has been chosen from available public datasets. SVT dataset is a word-spotting dataset and contains annotation of a few labeled text instances and ignores the rest of the text instances for tagging. Hence, this can be considered a good choice for the analysis of the missing annotation dataset. For training, we have considered the training set (100 images) of the SVT, whereas, for the testing, we rely on the fully annotated version of the testing set (250 images) by [20].

B. Performance Metrics

Suppose the ground truth bounding box for a dataset is B_g and the set of predicted bounding boxes by text detection approach is B_p. For text detection, traditionally, the evaluation protocol relies on the Intersection-over-Union (IoU) between B_g and B_p. The well-accepted text detection evaluation protocol comprises precision (P), recall (R), and f-measure (FM). Similar to the PASCAL VOC challenge protocol, in the text detection, the IoU threshold is taken as 0.5.

C. Experimental Setup

The experimental validation of the proposed approach has been conducted on the DGX machine having v100 GPU with 32 GB GPU memory. The implementation has been done on PyTorch [48] deep learning library.

TABLE I: Detection performance of the proposed method (refined segmentation based detection) and variant (text-biased segmentation based detection) for ICDAR 2015 dataset (with various drop rates) and SVT dataset. The best results are boldfaced. [P: Precision, R: Recall, FM: F-Measure]

Drop	Text-biased Segmentation			Refined Segmentation		
	P	**R**	**FM**	**P**	**R**	**FM**
ICDAR 2015						
0.6	0.865	0.722	0.787	0.889	0.697	0.781
0.4	0.804	0.806	0.805	0.840	0.801	0.820
0.2	0.828	0.820	0.824	0.861	0.812	0.836
0.0	0.829	0.836	0.832	0.859	0.827	0.843
SVT						
–	0.596	0.435	0.503	0.688	0.424	0.525

Training : For the training purpose, the starting layers of the proposed architecture, which are the same as the EAST method, have been initialized with the EAST architecture's pre-trained weights obtained with training on the synthetic dataset. The rest of the layers has been initialized with the He initialization approach [49]. The model has been trained using the Adam optimizer [50]. To avoid the over-fitting of the proposed model, we have used the L2 regularization with the factor of 0.0005. We fine-tuned the proposed model for 1000 epochs with a batch size of 12. The learning rate was initialized by 10^{-4}, and the learning rate decays at epoch 250 and 750 by 0.1. The runtime augmentation has been performed that includes rotation, scaling, and random cropping (512) of the image.

Inference : The inference of the proposed method has used the box regression branch and the refined segmentation branch in the text region refinement module. The bounding box generated from the regression branch has used the segmentation generated by the refined segmentation as a box score. Finally, the locality-aware non-maximal suppression has been used as a post-processing step.

D. Detection Performance

The detection performance evaluation has been performed on the generated ICDAR 2015 dataset having drop rates of 0.0, 0.2, 0.4, and 0.6 and on the publicly available SVT dataset. The proposed model has produced two segmentation maps. The first corresponds to the segmentation due to the hybrid loss, and the segmentation is text-biased, whereas the second is the refined segmentation. The evaluation has been conducted using both segmentations, and the results are reported in Table I. From Table I, it has been observed that the recall of the text-biased segmentation for detection is better than the refined segmentation. It proves that the text-biased segmentation gives more weightage to the text regions, hence, produce higher recall values than the refined segmentation. The refined segmentation helps to improve the precision with respect to the text-biased segmentation. As a result, f-measure is enhanced as compared with the text-biased segmentation. However, in the case of 0.6 drop of the ICDAR 2015 dataset, the decrease in recall of the refined segmentation with respect to text-biased segmentation is large than the other drop rates; the resulting

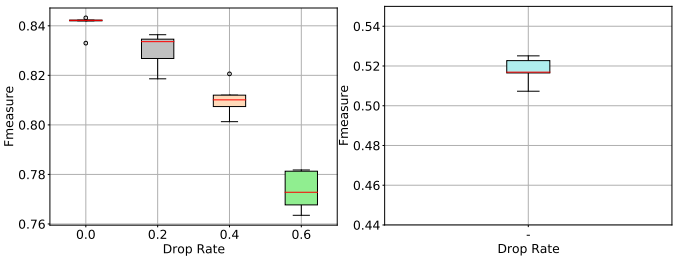


Fig. 5: Box plot to show the variance in the performance of the proposed method if the same training has been conducted five times. The median is shown by the red color line. (a) Box plot for generated (ICDAR 2015) datasets, (b) Box plot for public (SVT) dataset.

f-measure from the refined segmentation goes marginal lower than the text-biased segmentation's f-measure. In the other cases, refined segmentation has uplifted the precision with a marginal drop in the recall. Hence, the proposed method with refined segmentation has shown its strength to handle the partially annotated training data for generating good detection results.

E. Robustness Analysis

Run-time data augmentation has been used as an integral part of the proposed method. The augmentation sequence is not fixed in the run-time augmentation methods. As a consequence, augmented data may include less ambiguous pixels during the training phase. Hence, a single training might be lucky enough to get good results. Therefore, to show the proposed method's robustness, the same model has been trained five times. The performance of the proposed method is demonstrated in Fig. 5. For ICDAR 2015, the method's performance is most stable in the case of zero drop rate, and it is maximum in the case of the 0.6 drop case. The median is moving downside in the interquartile range of the box. The mean f-measure for generated dataset ICDAR 2015 with the drop rates of 0.6, 0.4, 0.2, and 0.0 are 0.7734, 0.8102, 0.8300, and 0.8405, respectively. For the SVT dataset, the mean f-measure is 0.5177. For further analysis, if the text-biased segmentation branch has used a scoring for bounding boxes. The mean f-measure from 5 runs for generated dataset ICDAR 2015 with the drop rates of 0.6, 0.4, 0.2, and 0.0 are 0.7812, 0.8066, 0.8183, and 0.8269, respectively. Whereas for the SVT dataset, the mean f-measure is 0.4934. However, the average f-measure of refined segmentation is better than text-biased segmentation for the SVT and ICDAR 2015 (with the drop rate of 0.4, 0.2, and 0.0). Hence, the refined segmentation is preferred over the text-biased segmentation in the proposed method during inference. This analysis shows the robustness of the proposed method for both publicly available as well as generated datasets.

F. Qualitative Analysis

The qualitative results of the baseline approach and the proposed approach are shown in Fig. 6. For generated dataset ICDAR 2015, Fig. 6 shows that in the baseline method, the boxes which appear on a lower drop rate disappear as the drop rate increases. In comparison, the proposed approach helps to retain the boxes even at the higher drop rate. It shows the robustness of the proposed method to retain the boxes even with the higher drop rate during the training phase. For the publicly available SVT dataset, the proposed method also shows better results than the baseline approach.

G. Comparative Analysis

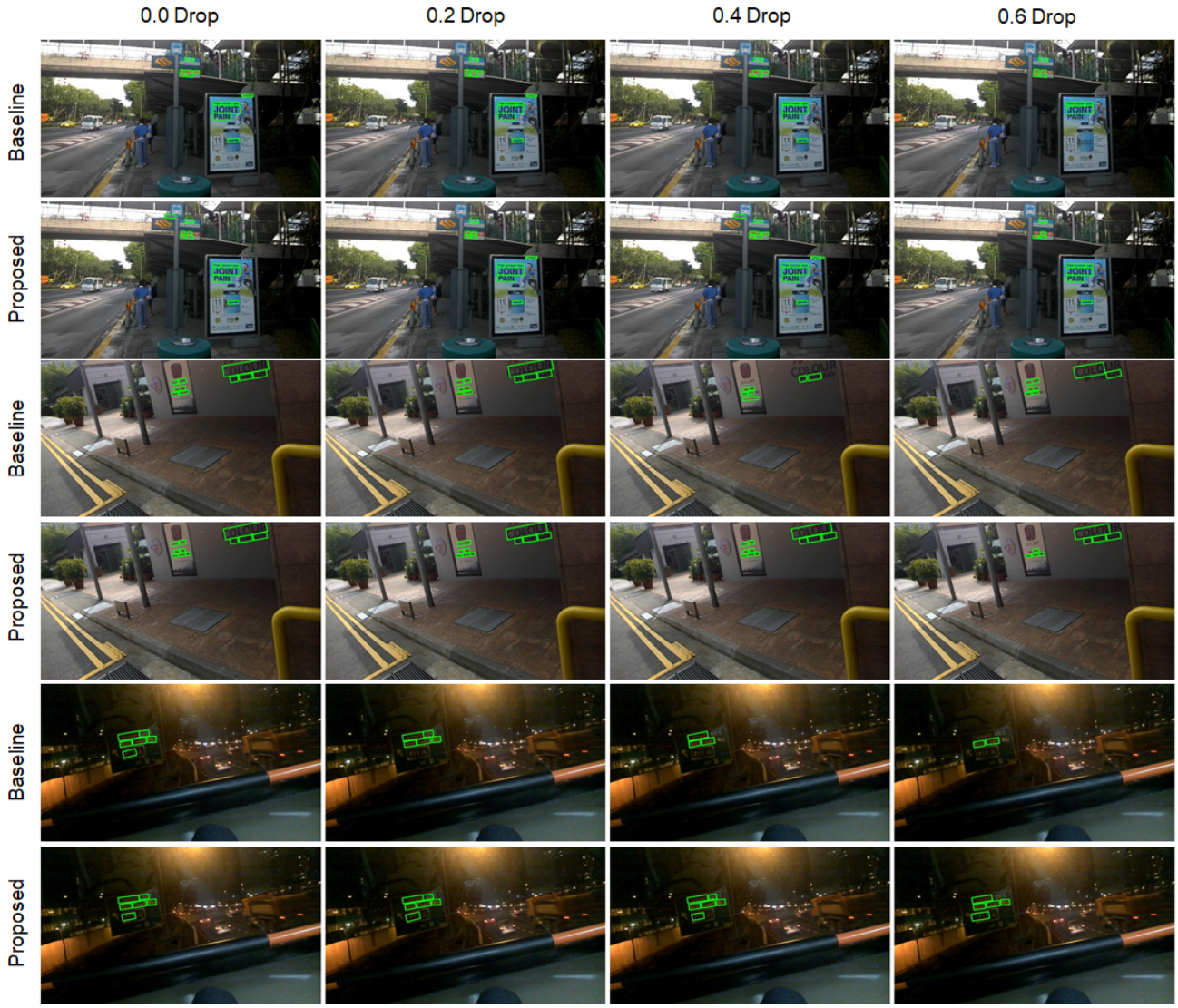
For the proper comparison of the work, single-stage text detectors have been chosen. The chosen methods for comparison are DB [21][1], EAST [26] [2],and QB [20]. A comparison among EAST, DB, QB, baseline and the proposed method has been reported in Table II. Table II shows that the proposed method outperforms the baseline approach from a significant margin in all drop rates for both generated (ICDAR 2015) and public dataset (SVT). The proposed method outperforms the EAST, DB and QB for the public SVT dataset as well as for ICDAR 2015 at drop rates of 0.2, 0.4, and 0.6. It is found that the proposed method is marginally inferior to the DB, EAST and QB only at the drop rate of 0.0 for the ICDAR 2015. From the comparison, it has been concluded that the text region refinement method is effective in handling the partially annotated training data for scene text detection.

In the baseline approach, the ground truth text regions excluded MSERs are treated as don't care regions for the model's training. Hence, it leaves a vast amount of background pixels with similar properties as text pixels. These regions ignore a huge number of background regions during training. It may lead to a low value of the recall. However, in the text region refinement method, the text potential regions are obtained by applying a more noise-tolerant loss. It predicts the text-probable regions, and using these regions as don't care regions is better than considering ground truth text regions excluded MSERs as don't care regions during training. The results show that this change in the baseline improves the results by a significant margin.

H. Ablation Study

The proposed method is based on the two-level text region refinement using pseudo-label generation for the partially annotated training dataset. For the ablation study, we have shown the impact of the text region refinement. The importance of text region refinement is shown by investigating the impact of adding refined segmentation to text-biased segmentation, i.e., used network with refinement (text-biased + refined segmentation) and without refinement (only text-biased segmentation), considering the rest of the setting same as the proposed approach. The performance has been mentioned in Table III. From Table III and II, it has been observed that the separate training with only hybrid loss also shows a vast improvement as compared to the baseline approach, which shows the impact of the hybrid loss function. From Table III, it also has been observed that the model without refinement block achieved

[1]https://github.com/MhLiao/DB
[2]https://github.com/SakuraRiven/EAST



(a)

(b)

Fig. 6: (a) ICDAR 2015 results are compiled in four columns. From left to right columns the drop rates are 0.0, 0.2, 0.4, and 0.6. (b) SVT results.

TABLE II: Comparative analysis of the proposed text region refinement method with other methods for partially annotated training datasets. The best results are boldfaced.

Dataset	DR	Model	P	R	FM
ICDAR 2015	0.6	EAST	0.901	0.267	0.412
		DB	0.910	0.293	0.443
		QB	0.867	0.297	0.442
		Baseline	0.856	0.453	0.592
		Proposed	0.889	0.697	**0.781**
	0.4	EAST	0.922	0.461	0.615
		DB	0.911	0.439	0.592
		QB	0.912	0.529	0.670
		Baseline	0.881	0.673	0.763
		Proposed	0.840	0.801	**0.820**
	0.2	EAST	0.892	0.700	0.785
		DB	0.914	0.624	0.742
		QB	0.857	0.731	0.789
		Baseline	0.838	0.769	0.802
		Proposed	0.861	0.812	**0.836**
	0.0	EAST	0.872	0.814	0.842
		DB	0.892	0.824	**0.856**
		QB	0.818	0.887	0.851
		Baseline	0.781	0.844	0.811
		Proposed	0.859	0.827	0.843
SVT	–	EAST	0.645	0.342	0.447
		DB	0.860	0.353	0.501
		QB	0.841	0.348	0.492
		Baseline	0.636	0.367	0.465
		Proposed	0.688	0.424	**0.525**

TABLE III: Ablation study to show the impact of the refinement module.

Dataset	DR	Refinement	P	R	FM
ICDAR 2015	0.6	✓	0.889	0.697	**0.781**
		✗	0.876	0.703	0.780
	0.4	✓	0.840	0.801	**0.820**
		✗	0.821	0.809	0.814
	0.2	✓	0.861	0.812	**0.836**
		✗	0.804	0.832	0.818
	0.0	✓	0.859	0.827	**0.843**
		✗	0.801	0.843	0.821
SVT	–	✓	0.688	0.424	**0.525**
		✗	0.625	0.427	0.507

better recall than the proposed refinement approach. It ensures the text-biased nature of hybrid loss. The addition of the refinement block increases the precision, but at the same time, a small penalty has been imposed on recall. However, for each dataset, the proposed method achieves better performance. Hence, this ablation justifies the impact and necessity of the hybrid loss and refinement phase.

I. Impact of pseudo-label

To understand the impact on pseudo-label in the training process, We have considered two metrics: the percentage of wrong gradient pixels (P_g) and the percentage of ignored pixels (P_i) for training data. The trade-off is shown in Fig. 7(a) and Fig. 7(b). We have considered the standard (DB, EAST, and QB), baseline, and proposed methods for this analysis. From Fig. 7(a), it has been found that as the drop rate increases, the number of pixels for the wrong gradient increases for all three categories of methods, but the rate

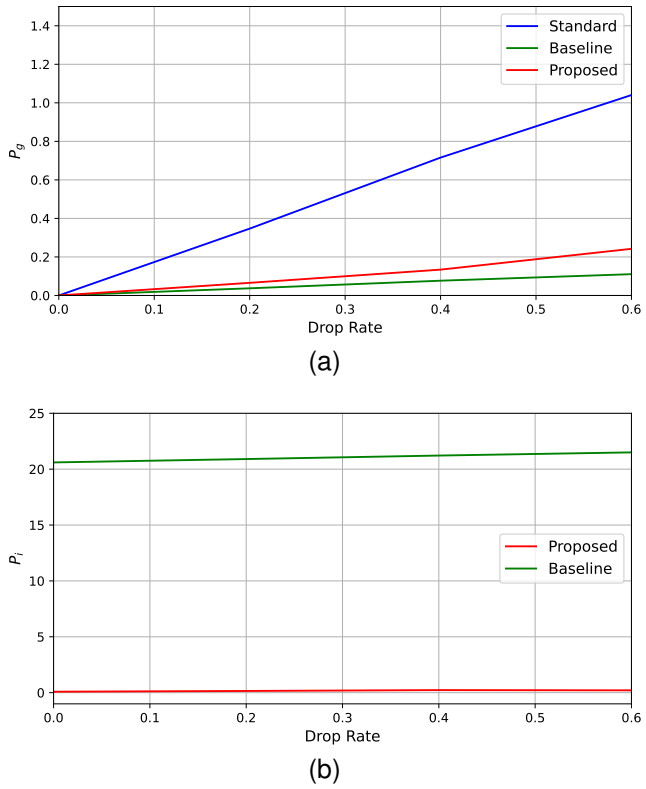


Fig. 7: (a) Trade-off between the drop rate and the percentage of wrong gradient pixels in training images, (b) trade-off between the drop rate and the percentage of ignored pixels in training images.

is slowest for the baseline method. From Fig. 7(a) and Fig. 7(b), it has been observed that the proposed method contains comparatively more wrong gradients than the baseline, but the ignored pixels are far less. Due to MSER regions drop, some background regions that possess the textual property have been dropped in the baseline. Whereas in the proposed method, pseudo-labeling helps to drop far fewer pixels with the penalty of a comparatively more percentage of wrong gradient pixels. In contrast, there is no inherent mechanism for dropping pixels in the standard methods. Therefore, the drop of image pixels may avoid learning the important clues similar to text strokes in training for baseline and proposed approaches. Hence, in the case of a completely annotated training dataset (i.e., drop rate 0.0), the baseline and proposed methods have dropped some of the background regions in training, leading to a drop in performance in the testing phase. The percentage of the ignored pixel of the proposed model is far less than the baseline due to pseudo-labeling, which helps achieve better performance than the baseline method. This finding is quantitatively supported by Table II.

J. Impact of threshold (T)

The proposed text region refinement consists of text-biased segmentation followed by refinement based on pseudo-labeling during the training phase. The pseudo-label generation depends on the choice of segmentation threshold (T) for the text-biased segmentation block. The impact of T has been tested

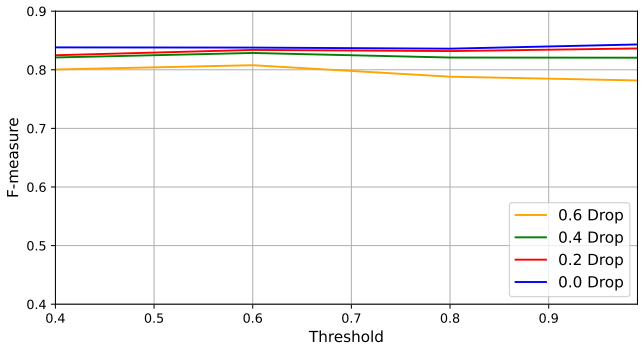


Fig. 8: Trade-off between the threshold (T) and the f-measure for various drop rates of the ICDAR 2015 dataset.

by training the various models for ICDAR 2015 dataset with different drop rates. The chosen thresholds for the analysis are 0.99, 0.8, 0.6, and 0.4. The line graph depicts the tradeoff between f-measure and thresholds in Fig. 8. From Fig. 8, at lower drop rates of 0.0 and 0.2, the higher threshold shows better performance. The number of ambiguous pixels (text pixels left unmarked during training) is already less at lower drop rates. Choosing a higher threshold may drop (i.e., don't care regions for refinement phase) an unnecessarily large number of pixels during training. But it may also decrease the performance for a higher threshold as compared to a lower threshold. The number of ambiguous pixels increases for the higher drop rate, and a higher threshold is suited to avoid the wrong label during training. Hence, at the drop rates of 0.4 and 0.6, the results improve as the threshold goes down. The method achieves best at the threshold of 0.6. From Fig. 8, it has been observed that the results are not deviating by a significant margin. There are marginal changes that have been found in Fig. 8. Hence, the proposed method also shows its robust nature against threshold (T) selection.

V. Conclusion and Future Work

This work investigates the dependence of text detection performance on the annotation quality during training. A text region refinement method for scene text detection has been proposed to handle missing annotations introduced due to the partial annotation of training data. The proposed method has significantly improved over the baseline approach and existing text detectors. In the training phase, ignoring certain regions does not let the network classify the background regions similar to text as a background. It improves the results of the baseline approach, but at the same time, it also limits the performance of the proposed method. Although the proposed method is validated in English, it can also be applied to other languages. The proposed method is specially designed to handle binary cases, limiting its application under the multi-script scenario. In the future, the method can be extended for multi-script text detection under the constraint of the partially annotated training data. The proposed method has assumed that the training image might contain partial annotation, and some background regions possessing the textual property have been dropped due to pseudo-labeling. It limits the performance of the proposed method while testing. In the future, we can add some extra images with no textual content during the training to include the background regions with textual properties and use this prior knowledge to skip the pseudo-label for these additional images.

References

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

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

[3] C. Zhang, W. Ding, G. Peng, F. Fu, and W. Wang, "Street view text recognition with deep learning for urban scene understanding in intelligent transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[4] S. Long, X. He, and C. Yao, "Scene text detection and recognition: The deep learning era," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 161–184, 2021.

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

[6] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 845–869, 2013.

[7] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Medical Image Analysis*, vol. 65, p. 101759, 2020.

[8] K. Wang and S. Belongie, "Word spotting in the wild," in *European Conference on Computer Vision*. Springer, 2010, pp. 591–604.

[9] F. Wang, L. Chen, C. Li, S. Huang, Y. Chen, C. Qian, and C. C. Loy, "The devil of face recognition is in the noise," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 765–780.

[10] X. Ma, Y. Wang, M. E. Houle, S. Zhou, S. Erfani, S. Xia, S. Wijewickrema, and J. Bailey, "Dimensionality-driven learning with noisy labels," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3355–3364.

[11] Y. Xu, L. Zhu, Y. Yang, and F. Wu, "Training robust object detectors from noisy category labels and imprecise bounding boxes," *IEEE Transactions on Image Processing*, vol. 30, pp. 5782–5792, 2021.

[12] L. Yang, F. Meng, H. Li, Q. Wu, and Q. Cheng, "Learning with noisy class labels for instance segmentation," in *European Conference on Computer Vision*, 2020, pp. 38–53.

[13] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1765–1773.

[14] J. Speth and E. M. Hand, "Automated label noise identification for facial attribute recognition." in *CVPR Workshops*, 2019, pp. 25–28.

[15] Z. Wu, N. Bodla, B. Singh, M. Najibi, R. Chellappa, and L. S. Davis, "Soft sampling for robust object detection," *arXiv preprint arXiv:1806.06986*, 2018.

[16] M. Xu, Y. Bai, B. Ghanem, B. Liu, Y. Gao, N. Guo, X. Ye, F. Wan, H. You, D. Fan *et al.*, "Missing labels in object detection." in *CVPR Workshops*, 2019.

[17] Y. Niitani, T. Akiba, T. Kerola, T. Ogawa, S. Sano, and S. Suzuki, "Sampling techniques for large-scale object detection from sparsely annotated objects," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6510–6518.

[18] N. Manwani and P. Sastry, "Noise tolerance under risk minimization," *IEEE Transactions on Cybernetics*, vol. 43, no. 3, pp. 1146–1151, 2013.

[19] G. Patrini, F. Nielsen, R. Nock, and M. Carioni, "Loss factorization, weakly supervised learning and label noise robustness," in *International Conference on Machine Learning*, 2016, pp. 708–717.

[20] P. Keserwani, A. Dhankhar, R. Saini, and P. P. Roy, "Quadbox: Quadrilateral bounding box based scene text detection using vector regression," *IEEE Access*, vol. 9, pp. 36 802–36 818, 2021.

[21] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-time scene text detection with differentiable binarization," in *Proc. AAAI*, 2020.

[22] Y. Cai, C. Liu, P. Cheng, D. Du, L. Zhang, W. Wang, and Q. Ye, "Scale-residual learning network for scene text detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[23] P. Cheng, Y. Cai, and W. Wang, "A direct regression scene text detector with position-sensitive segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 4171–4181, 2019.

[24] P. Lyu, M. Liao, C. Yao, W. Wu, and X. Bai, "Mask Textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 67–83.

[25] K. Raghunandan, P. Shivakumara, S. Roy, G. H. Kumar, U. Pal, and T. Lu, "Multi-script-oriented text detection and recognition in video/scene/born digital images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 4, pp. 1145–1162, 2018.

[26] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: An efficient and accurate scene text detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5551–5560.

[27] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Deep direct regression for multi-oriented scene text detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 745–753.

[28] P. Keserwani and P. P. Roy, "Text region conditional generative adversarial network for text concealment in the wild," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 3152–3163, 2022.

[29] Y. Sun, J. Liu, W. Liu, J. Han, E. Ding, and J. Liu, "Chinese street view text: Large-scale chinese text reading with partially supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9086–9095.

[30] D. Zhang, J. Han, G. Cheng, and M.-H. Yang, "Weakly supervised object localization and detection: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[31] M. Luo, X. Chang, L. Nie, Y. Yang, A. G. Hauptmann, and Q. Zheng, "An adaptive semisupervised feature analysis for video semantic recognition," *IEEE Transactions on Cybernetics*, vol. 48, no. 2, pp. 648–660, 2017.

[32] K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, and F. Nie, "A semisupervised recurrent convolutional attention model for human activity recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 5, pp. 1747–1756, 2019.

[33] Y. Zhang, M. Ding, Y. Bai, M. Xu, and B. Ghanem, "Beyond weakly supervised: Pseudo ground truths mining for missing bounding-boxes object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 983–997, 2019.

[34] H. Zhang, F. Chen, Z. Shen, Q. Hao, C. Zhu, and M. Savvides, "Solving missing-annotation object detection with background recalibration loss," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 1888–1892.

[35] Y. Yang, K. J. Liang, and L. Carin, "Object detection as a positive-unlabeled problem," *arXiv preprint arXiv:2002.04672*, 2020.

[36] T. Wang, T. Yang, J. Cao, and X. Zhang, "Co-mining: Self-supervised learning for sparsely annotated object detection," *arXiv preprint arXiv:2012.01950*, 2020.

[37] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[38] Y. Yang, Y. Zhuang, and Y. Pan, "Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies," *Frontiers of Information Technology & Electronic Engineering*, vol. 22, no. 12, pp. 1551–1558, 2021.

[39] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and vision computing*, vol. 22, no. 10, pp. 761–767, 2004.

[40] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Robust text detection in natural scene images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 970–983, 2013.

[41] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced MSER trees," in *European Conference on Computer Vision*. Springer, 2014, pp. 497–511.

[42] L. Gómez and D. Karatzas, "Textproposals: a text-specific selective search algorithm for word spotting in the wild," *Pattern Recognition*, vol. 70, pp. 60–74, 2017.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[44] A. Ghosh, H. Kumar, and P. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[45] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," in *International Conference on Medical Imaging with Deep Learning*. PMLR, 2019, pp. 285–296.

[46] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, and A. L. Martel, "Loss odyssey in medical image segmentation," *Medical Image Analysis*, vol. 71, p. 102035, 2021.

[47] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu *et al.*, "ICDAR 2015 competition on robust reading," in *13th International Conference on Document Analysis and Recognition*. IEEE, 2015, pp. 1156–1160.

[48] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Prateek Keserwani received the B.Sc., and M.Sc.(CS) from the University of Allahabad, Allahabad, India, in 2008, and 2010 respectively. He received M.Tech from University of Allahabad in 2015. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, Roorkee, India. His current research interests includes computer vision, brain signal analysis, and deep learning.

Rajkumar Saini received the Ph.D. degree from the Department of Computer Science and Engineering, IIT Roorkee, Roorkee, India. He is currently a Postdoctoral Researcher with the EISLAB Machine Learning, Luleå University of Technology, Sweden. His research interests include computer vision, machine learning, pattern recognition, human–computer interface, brain signal analysis, and digital image processing.

Marcus Liwicki received the M.S. degree in computer science from the Free University of Berlin, Germany, in 2004, and the Ph.D. degree from the University of Bern, Switzerland, in 2007. He worked as a Senior Researcher and a Lecturer with the German Research Center for Artificial Intelligence (DFKI). He is currently a Professor and the Head of machine learning subject with LTU University. He is a member of the International Association for Pattern Recognition (IAPR). He is a Program Committee Member of the IEEE ISM Workshop on Multimedia Technologies for E-learning. His research interests include knowledge management, semantic desktop, electronic pen-input devices, online and offline handwriting recognition, and document analysis.

Partha Pratim Roy is presently working as an Associate Professor in the Department of Computer Science and Engineering, Indian Institute of Technology (IIT), Roorkee. He received his Masters in 2006 and Ph.D. in 2010 from Universitat Autonoma de Barcelona, Spain. He did postdoctoral stays in France and Canada from 2010 to 2013. Dr. Roy gathered industrial experience while working in TCS and Samsung. In Samsung, he was a part-leader of the Computer Vision research team. He is the recipient of the "Best Student Paper" awarded by the International Conference on Document Analysis and Recognition, 2009, Spain. He has published more than 250 research papers in various international journals, conference proceedings. He has co-organized several international conferences and workshops, has been a member of the Program Committee of a number of international conferences and acts as a reviewer for many journals in the field. His research interests include Pattern Recognition, Document Image Processing, Biometrics, and Human-Computer Interaction. He is presently serving as Associate Editor of IET Image Processing, IET Biometrics, IEICE Transactions on Information and Systems, Springer Nature Computer Science.