Robust Scene Text Detection via Learnable Scene Transformations

Yuheng Cao 220171326@SEU.EDU.CN
School of Electronic Science and Engineering, Southeast University, Nanjing, China

Mengjie Zhou MENGJIE.ZHOU@BRISTOL.AC.UK
Department of Computer Science, University of Bristol, Bristol, UK

Jie Chen* JIECHEN01@PKU.EDU.CN
Peking University, Beijing, China

Editors: Emtiyaz Khan and Mehmet Gönen

Abstract

Scene text detection based on deep neural networks has been extensively studied in the last few years. However, the task of detecting texts in complex scenes such as bad weather and image distortions has not received sufficient attentions in existing works, which is crucial for real-world applications such as text translation, autonomous driving, etc. In this paper, we propose a novel strategy to automatically search for the effective scene transformation polices to augment images in the training phase. In addition, we build a new dataset, Robust-Text, to evaluate the robustness of text detection methods in real complex scenes. Experiments conducted on the ICDAR2015, MSRA-TD500 and Robust-Text datasets demonstrate that our method can effectively improve the robustness of text detectors in complex scenes.

Keywords: Scene Text Detection; Scene Transformation; Robustness Enhancements

1. Introduction

Scene text detection is widely used in image and video retrieval, autopilot and text translation, and has received intensive attention from researchers in areas of Computer Vision and Pattern Recognition. With the help of deep learning techniques, a great progress has been made in scene text detection. However, current deep learning based text detectors can be easily fooled in some complex scenes. As shown in Figure 1, when encountering object occlusion, bad weather or geometric distortions, existing state-of-the-art scene text detectors may still miss a large amount of texts, even though these texts are clearly visible to human eyes.

There exist some works Bahnsen et al. (2018); Mukherjee et al. (2018) that focus on modeling one single environmental condition, including removing reflection, fog, rain or snow. However, a variety of complex scenes exist in the wild, it is hard for the detector to know in advance which kind of conditions the target image belongs to. Moreover, if we design a model for each complex scene to eliminate the interference from the external environment, the entire detection process could be quite complicated and inefficient. There also exist some other works Vasiljevic et al. (2016); Zheng et al. (2016) that try to solve this problem by fine-tuning their models on blurred data. However, directly fine-tuning on blurred images may lead to underfitting, which significantly affects the performance of models.

* Indicates the corresponding author.

(a) Ground truth (b) EAST (c) SegLink (d) Mask R-CNN (e) Our method

Figure 1: Detection results of different methods in real natural scenes. (a) Ground truth. (b) Detection results of EAST. (c) Detection results of SegLink. (d) Detection results of MASK R-CNN. (e) Detection results of EAST with our method.

Based on the semantic segmentation maps and the properties of images from indoor or outdoor scenes, we design nine types of scene transformations including multi-perspective, reflection, fence, motion blur, fog, snow, glare, defocus blur and brightness. Furthermore, we adopt a strategy to automatically search for the transformation policies to augment the training images. To the best of our knowledge, this is the first work to improve the robustness of scene text detectors in various kinds of complex scenes. Moreover, a dataset named as Robust-Text containing 100 validation images and 400 test images from indoor and outdoor scenes is constructed to evaluate the robustness of scene text detectors in real complex scenes.

Major contributions of this paper can be summarized as follows:

- In order to simulate complex scenes, we design nine types of scene transformations and adopt a strategy to automatically search for optimal transformation policies to augment images in the training phase.
- A new dataset, Robust-Text, is proposed to evaluate the robustness of scene text detectors in real natural scenes.
- The learnable scene transformation technique can effectively improve the robustness of text detectors in complex scenes.

2. Related Work

2.1. Scene Text Detection

Scene text detection and recognition have been extensively studied for a long time. Before the prevalence of deep learning, a large number of conventional methods, such as Stroke Width Trans-

form (SWT) Epshtein et al. (2010) and Maximally Stable Extremal Regions (MSER) Neumann and Matas (2010), detect scene texts relying on manually designed features. Recently, with the help of deep neural networks, modern methods, which can be coarsely divided into segmentation-based Yao et al. (2016); Zhang et al. (2016); Xu et al. (2019); Wu and Natarajan (2017); Polzounov et al. (2017); Xue et al. (2018) and regression-based Tian et al. (2017, 2015); Liao et al. (2017); Ma et al. (2018); Lyu et al. (2018), markedly outperform the conventional methods.

However, complex scenes such as bad weather can easily fool deep learning based methods, which has a huge negative impact on detecting texts in real-world applications. Moreover, existing datasets contain few images in different complex scenes for text detection and collecting a large amount of images in complex scenes is difficult. So improving the robustness of scene text detectors in a simple and effective way becomes increasingly important.

2.2. Robustness Studies

In recent years, the frangibility of deep learning based methods to common corruptions has received more and more attention from researchers. Geirhos et al. Geirhos et al. (2018) compared the generalization capabilities of humans and deep neural networks on object recognition under a broad range of corruption types. Different corrupted datasets for object and traffic sign recognition were first proposed by Temel and AlRegib Temel and AlRegib (2018). Latter, Hendrycks and Dietterich Hendrycks and Dietterich (2019) established benchmarks consisting of the "IMAGENET-C" and "IMAGENET-P" datasets to evaluate the robustness of recognition models when affected by corruptions. Michaelis et al. Michaelis et al. (2019) also proposed three benchmarks composed of the "Pascal-C", "Coco-C" and "Cityscapes-C" datasets to assess the performance of object detection models when the image quality degrades. Actually, there are many other corruptions such as reflection and glare that often appear in the scenes for text detection, but they are not considered in above-mentioned benchmarks.

Existing methods designed to alleviate the performance degradation can be roughly classified into two categories: preprocessing the input data by removing corruption and fine-tuning models on corrupted data. For example, for the first category, based on CNN, Mukherjee et al. Mukherjee et al. (2018) proposed NVDeHazenet to restore the quality of images under rainy and foggy circumstances. However, these methods are designed for a specific category of corruption and when encountering interferences from other corruption categories, their performance can be greatly reduced. For the second category, Vasiljevic et al. Vasiljevic et al. (2016) fine-tuned a pre-trained model on corrupted images of a specific category, but the model cannot be generalized to other corruption types.

In order to simulate the common complex scenes, we design nine types of scene transformations including multi-perspective, reflection, fence, motion blur, fog, snow, glare, defocus blur and brightness to transform the training images.

2.3. Data Augmentation

Data augmentation has received intensive attention from researchers in areas of image recognition, object detection, etc. In the past few years, operations including random cropping, scaling, flipping, rotation and color transformation are commonly performed on benchmark datasets to generate the augmentation examples. Furthermore, there are some methods such as Cutout DeVries and Taylor (2017), Mixup Zhang et al. (2017) and CutMix Yun et al. (2019), which achieve promising results on

image recognition tasks by randomly replacing or masking out the image patches. However, these methods need to set proper hyper-parameters for specific datasets based on domain knowledge.

Drawing inspiration from the neural architecture search (NAS) algorithms, some current methods automatically search for the effective augmentation polices for corresponding datasets and models. Based on the Reinforcement learning (RL) search strategy, AutoAugment Cubuk et al. (2019) trains a RNN controller to predict the policies for augmenting images. PBA Ho et al. (2019) uses a population based training to generate the augmentation policies. Fast AutoAugment Lim et al. (2019) employs the Bayesian optimization to search for the optimal policies.

Based on the scene transformations mentioned above, we design the search space. Inspired by Fast AutoAugment Lim et al. (2019), we treat the task of finding optimal transformation policies as a density matching problem and apply the Bayesian optimization to learn the policies.

3. Method Description

3.1. Overview

The text detection models trained on existing datasets perform well on their corresponding test sets. However, the detection performance of these models will be significantly deteriorated when encountering widely varying indoor and outdoor scenes such as occlusion, reflection, fog, snow, motion blur and so on. Moreover, there is currently no dataset specifically designed for text detection in such complex scenes as mentioned above. Therefore, we adopt a strategy to automatically search for scene transformation policies to pre-process images on existing datasets to solve the problem of detecting texts in complicated scenes.

The pipeline of our scene transformation framework is illustrated in Figure 2. First, based on the target images, we trained a scene text detector. Next, the source image is segmented into different regions based on semantic information. Meanwhile, a discriminator predicts whether the image is from indoor or outdoor scenes. If the image comes from indoor scenes, we randomly select the indoor image materials from the image repository, otherwise the outdoor image materials are chosen. Then, we randomly select a sub-policy including two consecutive operations to transform the source image. Furthermore, we explore the policies via the Bayesian optimization method by minimizing the expected loss of the trained scene text detector on augmented images. Finally, by using the obtained polices, we perform scene transformations on the input images during the training phase.

3.2. Semantic Coherence

When performing scene transformations on an image, we need to consider the semantic information of each region in the image. For example, reflections typically occur on the surface of objects such as windows instead of grass, sky, etc. With the help of acquired semantic information, the transformed images are more semantically reasonable.

Based on the existing datasets for scene text detection, we trained a classifier to identify whether the input image is from indoor or outdoor scenes. The ground truth is manually labeled by us. To further understand the visual scene of images at pixel level, we perform semantic segmentation on images. There are a large number of objects labeled at pixel level in the COCO 2018 panoptic segmentation dataset Kirillov et al. (2019) which contains almost all common objects in natural scenes. Based on the dataset, we adopt the method proposed in Xiong et al. (2019) to automatically

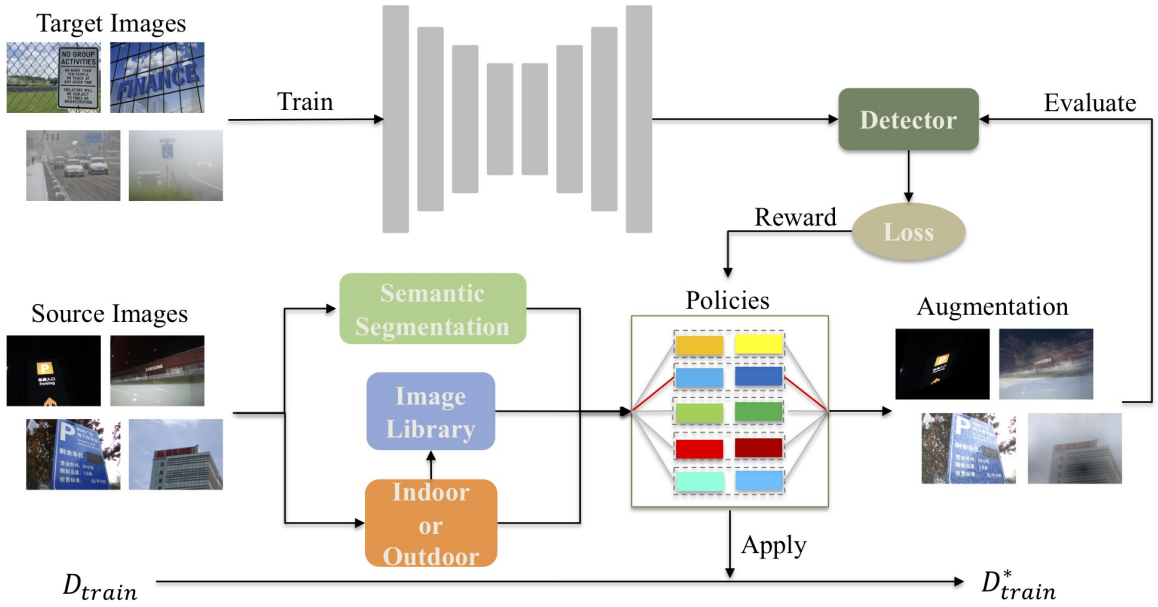

Figure 2: The pipeline of the proposed scene transformation framework. First, based on the target images, we train the scene text detector. Next, by using the semantic information and properties of indoor or outdoor, we randomly select a sub-policy including two consecutive operations to transform the source images. Furthermore, we explore the transformation policies by minimizing the expected loss of the scene text detector on the augmented images.

annotate all semantic classes in images. As illustrated in Figure 2, the obtained scene and semantic information is conducive to performing scene transformations on sensible regions. The details of using the obtained information in each type of scene will be explained in section 3.3.

3.3. Scene Transformations

When performing scene transformations, we need to distinguish between indoor and outdoor scenes due to their different properties. For example, there is no snow or fog indoor. In addition, we use a continuous value of magnitude m at $[0, 1]$ to control the severity levels of scene transformations. Note that some transformations (e.g., fence, glare) do not have the magnitude. As shown in Figure 3, to simulate complex scenes, we design nine types of scene transformations as follows.

Multi-perspective means taking pictures from multiple perspectives. Images in most existing datasets are taken from the front of texts, but it's commonly-seen for ordinary people to take pictures from many other perspectives in real situations. Therefore, we transform the images from four perspectives consisting of top left, top right, bottom left and bottom right.

Reflection can appear when taking images through the glass, which causes visual interference to scene text detection. Based on the semantic segmentation map of the image, we first select some appropriate semantic areas. Then, according to whether the image is from indoor or outdoor scenes,

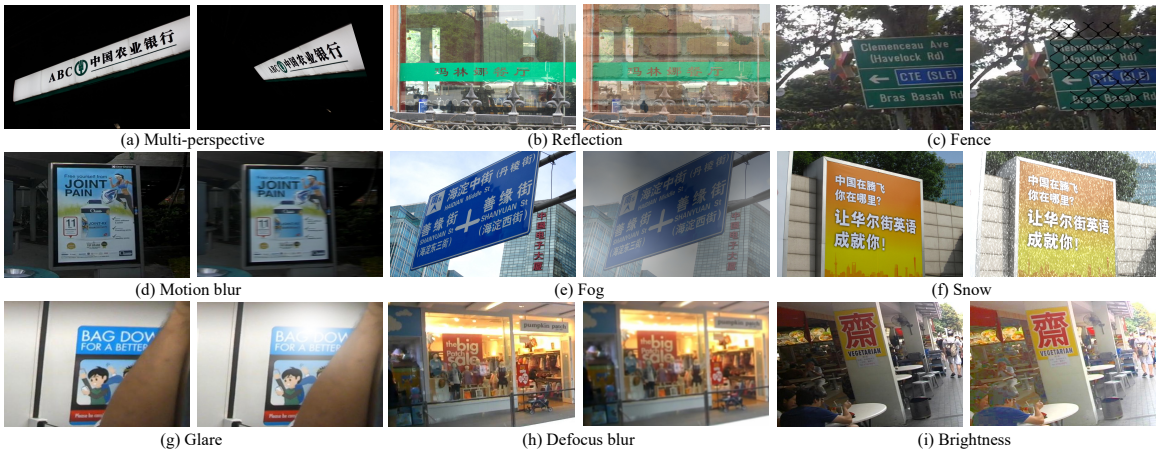

Figure 3: An illustration of the proposed nine types of scene transformations.

we mix some randomly-selected image materials in the corresponding category with the selected areas in the image by a weight of $r \in [0, 1]$, which can be used to control the magnitude of this transformation.

Fence is a kind of common objects occluding texts in natural scenes. We design three different types of fences and randomly select a fence to render the text regions in images with outdoor scenes.

Motion blur occurs when the camera moves quickly at the time of taking pictures. Given a radius, sigma and an angle in which the blur should occur, we perform motion blur on the entire image by using the motion-blur function of ImageMagick ImageMagick Studio (2008). The parameter sigma is used to set different severities of motion blur.

Fog can make objects unrecognizable, which has a huge negative impact on applications such as the autopilot in real natural scenes. We adopt the diamond-square algorithm Fournier et al. (1982) to render fog.

Snow generally appears in winter, which can cause large visual disturbances during scene text detection. We generate a blank image and randomly draw a certain number of white points on it. Furthermore, the motion-blur function of ImageMagick ImageMagick Studio (2008) is applied on the generated image to simulate the effect of snow falling. We generate snows in different severities by setting different numbers of white points and parameter sigma in the motion-blur function.

Glare may appear in some locations of scenes. Because of its high brightness, texts around it are very difficult to detect. To ensure the recognizability of scene texts, the strong light points are mainly placed at the border of texts.

Defocus blur refers to the situation where the image is out of focus. We implement this type of scene transformation through the gaussian blur function in OpenCV Bradski (2000).

Brightness of a scene changes with the external illumination of varying intensities. An environment that is too bright or too dark may seriously affect the detection of scene texts. Skimage Van der Walt et al. (2014), a popular Python image library, is used to implement different severities of brightness on images.

3.4. Searching for Transformation Policies

Search Space The proposed scene transformations (augmentation operations) mentioned above are employed to transform the input images. Following Fast AutoAugment Lim et al. (2019), our transformation policies P include several sub-policies where a sub-policy $s_i \in P$ consists of N consecutive operations $\{\overline{O}_n^{s_i}(x; p_n^{s_i}, m_n^{s_i}) : n = 1, ..., N\}$, where N is set to 2 in our experiments. Each operation $\overline{O}_n^{s_i}$ applied to an input image has two continuous parameters: the calling probability of applying this operation $p_n^{s_i}$ and the magnitude of this operation $m_n^{s_i}$:

$$\overline{O}_n^{s_i}(x; p_n^{s_i}, m_n^{s_i}) := \begin{cases} O(x; m_n^{s_i}) & : probability\ p_n^{s_i} \\ x & : probability\ 1 - p_n^{s_i}. \end{cases} \tag{1}$$

Therefore, a sub-policy s_i can be represented by a composition of operations as follows:

$$s_i = \overline{O}_N^{s_i}((..., \overline{O}_2^{s_i}((\overline{O}_1^{s_i}(x; p_1^{s_i}, m_1^{s_i})); p_2^{s_i}, m_2^{s_i})); p_N^{s_i}, m_N^{s_i}). \tag{2}$$

Search Strategy As shown in Figure 2, we search for the transformation policies by matching density between the source and target images. Based on the target images, we first train the scene text detection model. Furthermore, we explore the transformation policies by tuning calling probabilities and magnitudes of transformation operations for minimizing the expected loss of the trained model on augmented source images:

$$s_i^* = \arg\min_{s_i} Loss(\theta^* | s_i(D_s)), \tag{3}$$

where θ^* is the parameter trained on the target images D_t, D_s denotes the source images and s_i^* approximately minimizes the distance between the densities of D_t and $s_i(D_s)$ by minimizing the expected loss.

Implementation Based on the Bayesian optimization, we explore the desired transformation policies. Here, we present the implementation details of each step listed in Algorithm 1. In this paper, *Tune* Liaw et al. (2018), a scalable hyperparameter tuning library, is used for policy searching. We explore the the continuous values for the calling probability p and magnitude m for each operation. At the beginning, the values of p and m are uniformly sampled from [0, 1]. Then, HyperOpt in *Tune* modifies the values to minimize the $Loss$.

Algorithm 1 Searching for Scene Transformation Policies

Input: $(\theta, D_{source}, D_{target}, s, T, P)$
  $P \leftarrow \varnothing$
  Train θ on D_{target}
  for $t \in \{0, ..., T-1\}$ **do**
    $\mathbb{S} \leftarrow$ BayesOptim$(s, Loss(\theta | s(D_{source})))$
    $S_n \leftarrow$ Select top-n polices in $\mathbb{S}$
    $P \leftarrow P \cup S_n$
  end for
Output: P

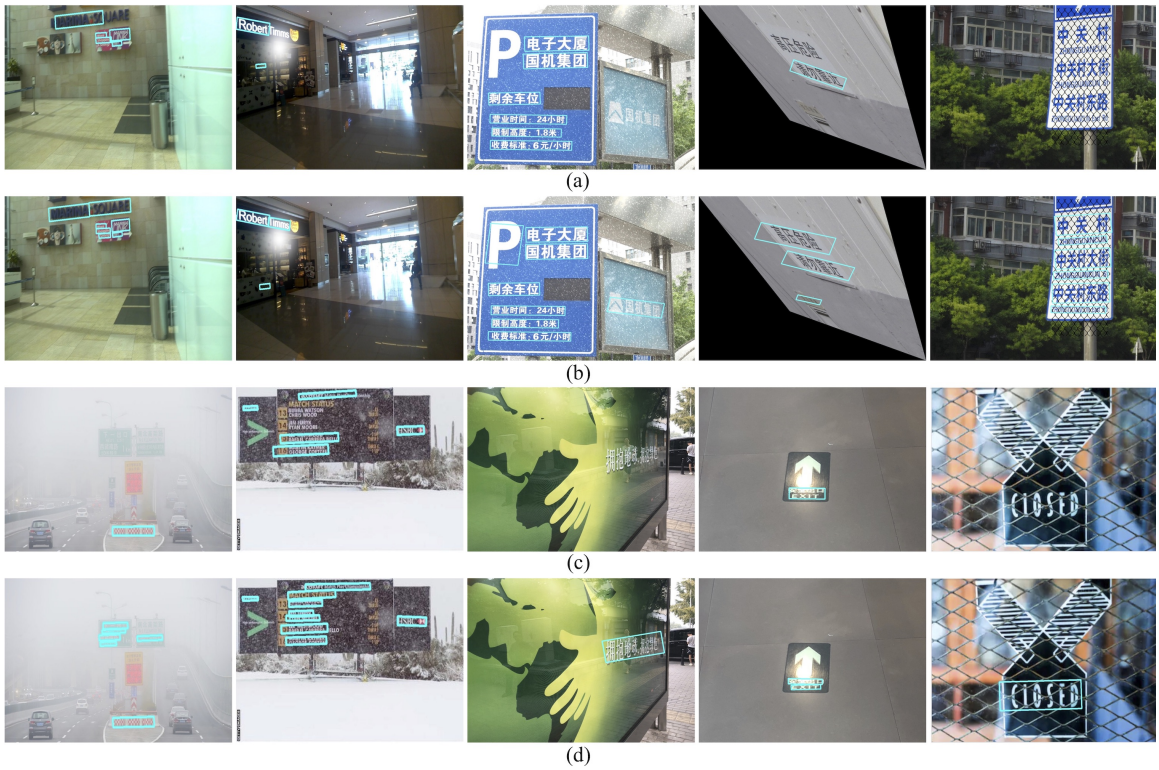

Figure 4: Some qualitative detection results on transformed and real natural images. (a) The detection results on transformed images of EAST trained on the original datasets. (b) The detection results on transformed images of EAST trained on the transformed images by using our learnable scene transformation technique. (c) The detection results on Robust-Text of EAST trained on MSRA-TD500. (d) The detection results on Robust-Text of EAST trained on the transformed MSRA-TD500 by using the proposed learnable scene transformation technique.

4. Robust-Text Dataset

There is currently no dataset specifically designed for detecting texts in complex scenes, but these scenes do exist in real world and pose great challenges for detection tasks. To evaluate the performance of models on detecting texts in complex scenes, we propose a dataset named Robust-Text containing 100 validation images and 400 test images, which include images manually collected from internet and our own pictures taken by a smart phone. These images are from both indoor and outdoor scenes, covering a variety of complex scenes such as snow, fog, multi-perspective, glare, reflection, etc. Some representative images of the proposed dataset are illustrated in Figure 4. Moreover, texts on those images may be in Chinese, English or the mixture of both. We manually label texts by using our own labeling tool and all texts in images are annotated at line level by using minimum bounding rectangles.

Table 1: Empirically chosen magnitudes for different types of scene transformations

Scene Types	Multi-perspective	Reflection	Fence	Motion blur	Fog	Snow	Glare	Defocus blur	Brightness
Magnitude	0.5	0.3	-	0.3, 0.5	0.3,0.5,0.8	0.3,0.5,0.8	-	0.5	0.3,0.5,0.8

5. Experiments

In this section, we conduct experiments to evaluate several state-of-the-art scene text detection methods in complex scenes and analyze their detection performance in each scene. Experimental results demonstrate that the proposed learnable scene transformation technique effectively improves the robustness of text detectors in complex scenes.

5.1. Datasets

ICDAR2015 and ICDAR2015-R Datasets ICDAR2105 Karatzas et al. (2015) was used in the Challenge 4 of 2015 Robust Reading Competition. This dataset contains 1000 images for training and 500 images for testing, which are captured without preparation by Google Glasses. The text instances are annotated as quadrilaterals at word level. ICDAR2015-R, including 1000 images for validation and 3259 images for testing, is generated by performing nine types of scene transformations on ICDAR2015. As shown in Table 1, to simulate the situations of real scenes, we empirically set the magnitude of each scene transformation for ICDAR2015-R.

MSRA-TD500 and MSRA-TD500-R Datasets MSRA-TD500 Yao et al. (2012) is a dataset that contains 300 training images and 200 testing images, which are taken from indoor and outdoor scenes. The text instances in images are annotated at line level. MSRA-TD500-R is acquired by performing nine types of scene transformations on MSRA-TD500 dataset. This dataset consists of 500 validation images and 1746 testing images. The magnitude settings of different scene transformations in MSRA-TD500-R are the same as those in ICDAR2015-R (see Table 1).

Robust-Text Dataset The dataset we built as mentioned in Section 4.

5.2. Scene Text Detection

Impact of Complex Scenes In order to assess the impact of complex scenes, we evaluate several existing text detection methods by training models on the training sets of ICDAR2015 and MSRA-TD500 and testing them on the test sets of ICDAR2015, ICDAR2015-R, MSRA-TD500 and MSRA-TD500-R.

As we can observe from results listed in Table 2, complex scenes have a huge negative impact on the performance of scene text detection methods. The performance of EAST Zhou et al. (2017) drops significantly by 18.0% and 10.5% in f-score on ICDAR2015-R and MSRA-TD500-R, respectively and PAN Wang et al. (2019) suffers a dramatic reduction in f-score by 19.1% and 17.3% on ICDAR2015-R and MSRA-TD500-R, respectively. We also evaluate CRAFT Baek et al. (2019) with VGG-16 as the backbone network, whose performance drops by 17.7% and 18.2% in f-score on ICDAR2015-R and MSRA-TD500-R, respectively. Furthermore, a recently-proposed method DB-ResNet-50 Liao et al. (2019) does not perform well in this condition, which suffers a reduction in f-score by 17.5% and 17.6% on ICDAR2015-R and MSRA-TD500-R, respectively.

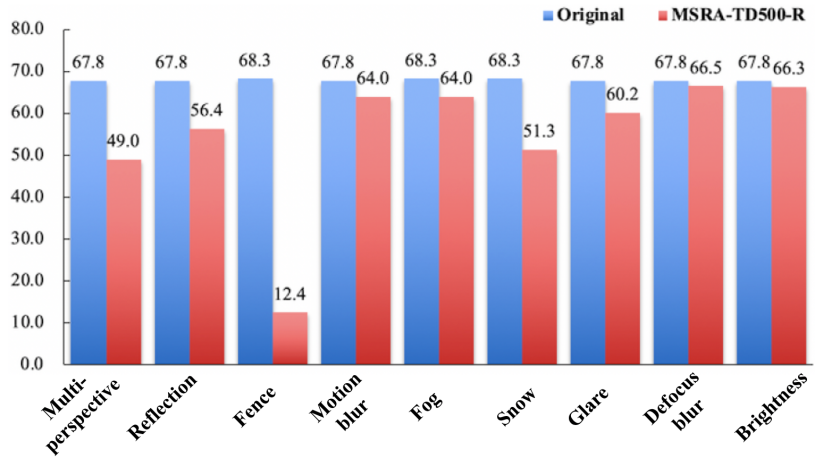

Figure 5: The comparison of detection performance of EAST in different scenes. The model is trained on the training set of MSRA-TD500. Bars colored in blue and red denote results obtained by testing the model on the test set of MSRA-TD500 and MSRA-TD500-R, respectively.

Furthermore, based on EAST Zhou et al. (2017), we conduct experiments on MSRA-TD500 and MSRA-TD500-R to analyze the impact of various scene transformation types on the detection results. As shown in Figure 5, fence has the greatest interference to detection results, which decreases the performance by 55.9% in f-score. In addition, multi-perspective, reflection, fog, snow and glare also greatly affect the detection results, while the interference caused by brightness, motion blur and defocus blur is relatively small.

Robustness Enhancements For the purpose of improving the robustness of scene text detection methods, we regard the validation sets of ICDAR2015-R and MSRA-TD500-R as the target images to search for the corresponding scene transformation policies. Based on the searched policies, we perform scene transformations on the training sets of ICDAR2015 and MSRA-TD500. Table 2 shows the quantitative results on the test sets of four datasets.

On the one hand, we analyze the effect of the proposed method on the test sets of original datasets. 'LST' stands for the proposed learnable scene transformation technique. From Table 2, we can observe that EAST [LST] outperforms the original EAST in f-score by 2.5% and 1.4% on ICDAR2015 and MSRA-TD500, respectively. Furthermore, PAN [LST] outperforms the original PAN in f-score by 1.0% and 0.9% on ICDAR2015 and MSRA-TD500, respectively. The other methods also get some improvements on the ICDAR2015 and MSRA-TD500.

On the other hand, we analyze the effect of the proposed method on the test sets of transformed datasets. The detection results of EAST [LST] just decrease by 9.9% and 1.6% on ICDAR2015-R and MSRA-TD500-R, respectively, while the performance of the original EAST dramatically drops by 18.0% and 10.5% in f-score on ICDAR2015-R and MSRA-TD500-R, respectively. In addition, the detection results of DB-ResNet-50 [LST] decrease by 9.8% and 2.5% on ICDAR2015-R and MSRA-TD500-R, respectively, while the performance of the original DB-ResNet-50 dramatically drops by 17.5% and 17.6% in f-score on ICDAR2015-R and MSRA-TD500-R, respectively.

Table 2: The detection results of several methods on ICDAR2015, ICDAR2015-R, MSRA-TD500 and MSRA-TD500-R. These methods are trained on the training images of ICDAR2015 and MSRA-TD500. 'Validation' means adding the validation images of ICDAR2015-R or MSRA-TD500-R based on the training images of ICDAR2015 or MSRA-TD500. 'LST' stands for the proposed learnable scene transformation technique.'R', 'P' and 'F' mean the 'recall', 'precision' and 'f-score', respectively.

	ICDAR2015			**ICDAR2015-R**			**MSRA-TD500**			**MSRA-TD500-R**		
Method	R	P	F	R	P	F	R	P	F	R	P	F
PAN Wang et al. (2019)	81.9	84.0	82.9	52.1	82.3	63.8	83.8	84.4	84.1	57.1	80.5	66.8
CRAFT Baek et al. (2019)	84.3	89.8	86.9	57.5	86.8	69.2	78.2	88.2	82.9	52.2	85.1	64.7
DB-ResNet-50 Liao et al. (2019)	80.3	**91.3**	85.4	54.9	**88.9**	67.9	80.8	**91.1**	85.6	55.7	87.2	68.0
EAST Zhou et al. (2017)	76.9	81.1	79.0	48.9	81.1	61.0	60.9	79.0	68.8	46.5	78.2	58.3
PAN Wang et al. (2019) + Validation	82.1	85.1	83.6	65.8	75.0	70.1	84.6	85.6	85.1	72.5	75.8	74.1
CRAFT Baek et al. (2019) + Validation	84.3	89.0	86.6	66.5	77.3	71.5	79.0	87.9	83.2	72.5	80.8	76.4
DB-ResNet-50 Liao et al. (2019) + Validation	82.2	91.3	86.5	74.2	68.6	71.3	82.7	89.8	86.1	72.9	81.8	77.1
EAST Zhou et al. (2017) + Validation	78.8	82.9	80.8	61.5	72.4	66.5	63.1	78.1	69.8	53.6	76.4	63.0
PAN Wang et al. (2019) + LST	82.9	84.9	83.9	68.9	78.3	73.3	**85.3**	84.7	85.0	79.3	82.1	80.7
CRAFT Baek et al. (2019) + LST	**84.6**	89.7	**87.1**	72.5	80.9	**76.5**	79.2	88.0	83.4	77.4	85.6	81.3
DB-ResNet-50 Liao et al. (2019) + LST	82.0	90.8	86.2	**79.1**	72.0	75.4	82.2	90.8	**86.3**	**80.1**	**87.8**	**83.8**
EAST Zhou et al. (2017) + LST	79.2	83.9	81.5	66.8	77.1	71.6	63.3	78.8	70.2	59.3	81.4	68.6

Furthermore, we add the validation images of '-R' dataset in the training phase to exclude the effects of complex training data. As shown in Table 2, LST outperforms simply adding the validation images of ICDAR2015-R or MSRA-TD500-R in the complex scenes.

From the experimental results mentioned above, we can observe that the proposed scene transformation technique can be conducive to improving the robustness of detectors in complex scenes. Moreover, the performance of detectors can also be slightly improved on the original dataset in simple scenes. Some samples of our detection results on ICDAR2015 and MSRA-TD500 are illustrated and compared in Figure 4.

Text Detection in Real Complex Scenes To further verify the effectiveness of our method in real natural scenes, based on some widely-used scene text detectors, we conduct several experiments on our Robust-Text dataset. Because the Robust-Text and MSRA-TD500 datasets both are taken from indoors and outdoors and annotated at text-line level, we adopt the training set of MSRA-TD500 for training and the test set of Robust-Text for testing. Based on the validation set of Robust-Text, we search for the corresponding transformation policies.

'RST' denotes the random scene transformation. As shown in Table 3, CRAFT [RST] surpasses CRAFT by 7.0% (56.1 vs 63.1) in f-score and EAST [RST] surpasses EAST by 4.6% (54.1 vs 58.7) in f-score. Observing the detection results of PAN [LST] and PAN, a conclusion can be made that the learnable scene transformation technique significantly improves the f-score of detection results by 9.1% (56.7 vs 65.8). Furthermore, DB-ResNet-50 [LST] outperforms DB-ResNet-50 [RST] in f-score by 5.3% (61.4 vs 66.7), which demonstrates that the learnable scene transformation is more effective than the random scene transformation. Moreover, DB-ResNet-50 [LST] outperforms DB-ResNet-50 [Validation] in f-score by 4.1% (62.6 vs 66.7), which verifies that 'LST' is better than

Table 3: Results of ablation studies for our method on the proposed Robust-Text dataset. EAST is adopted for evaluation in experiments. 'Validation' means adding the validation images of Robust-Text dataset based on the training images of MSRA-TD500. 'RST' and 'LST' denote the random scene transformation and the learnable scene transformation technique, respectively. 'R', 'P' and 'F' mean the 'recall', 'precision' and 'f-score', respectively.

Method	R	P	F
PAN Wang et al. (2019)	46.9	71.6	56.7
PAN Wang et al. (2019) + RST	58.9	67.0	62.6
PAN Wang et al. (2019) + Validation	57.6	69.8	63.1
PAN Wang et al. (2019) + LST	**60.2**	72.5	65.8
CRAFT Baek et al. (2019)	45.1	74.2	56.1
CRAFT Baek et al. (2019) + RST	56.5	71.5	63.1
CRAFT Baek et al. (2019) + Validation	55.8	73.7	63.5
CRAFT Baek et al. (2019) + LST	56.9	75.8	65.0
DB-ResNet-50 Liao et al. (2019)	45.2	77.7	57.2
DB-ResNet-50 Liao et al. (2019) + RST	53.1	72.8	61.4
DB-ResNet-50 Liao et al. (2019) + Validation	54.1	74.2	62.6
DB-ResNet-50 Liao et al. (2019) + LST	58.0	**78.4**	**66.7**
EAST Zhou et al. (2017)	44.1	70.0	54.1
EAST Zhou et al. (2017) + RST	54.1	64.2	58.7
EAST Zhou et al. (2017) + Validation	54.0	67.1	59.8
EAST Zhou et al. (2017) + LST	56.7	70.6	62.9

simply adding complex training data. Some examples of our detection results on Robust-Text are shown and compared in Figure 4.

Comparison with the Image Dehazing Method In this section, an experiment on detecting texts in foggy weather is conducted to compare our method with the image dehazing method DCPDN Zhang and Patel (2018). We compare the detection results on images dehazed by DCDPN of EAST trained on MSRA-TD500 with the detection results on original images of EAST trained on the transformed MSRA-TD500 by using our method with the learnable scene transformation technique. The qualitative results are shown in Figure 6, from which we can see that the detection performance of EAST by using our method (the second line in Figure 6) is comparable to that of EAST by removing the fog in advance (the first line in Figure 6), which demonstrates the effectiveness of our method. Moreover, our method can be applied to a variety of complex scenes, while the image dehazing method is only suitable for a single scene.

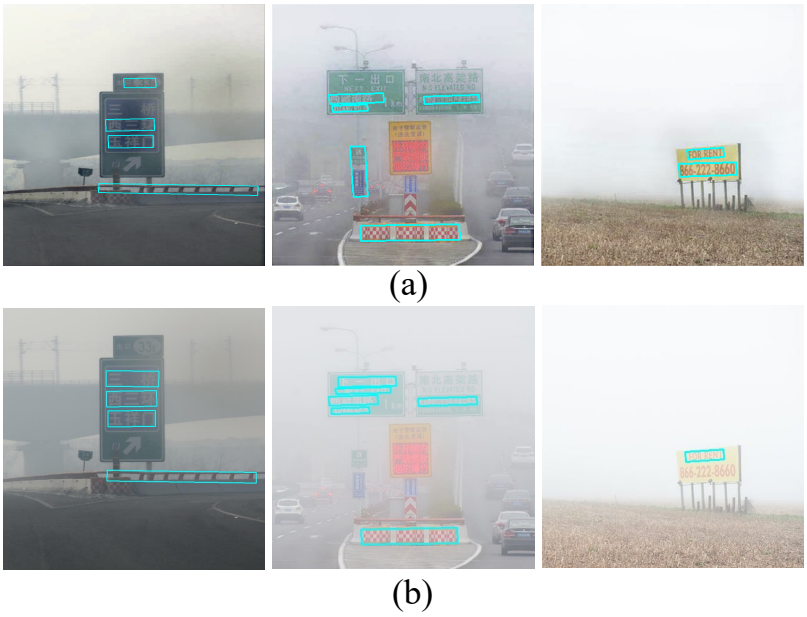

Figure 6: Comparison with the image dehazing method. (a) The detection results on images dehazed by DCDPN Zhang and Patel (2018) of EAST trained on MSRA-TD500. (b) The detection results on original images of EAST trained on the transformed MSRA-TD500 by using our method.

6. Conclusion

This paper presented a learnable scene transformation framework to effectively improve the robustness of text detectors in complex scenes. Specifically, we designed nine types of scene transformations and automatically search for the transformation policies to simulate complex scenes during the training phase. Moreover, we built a new dataset Robust-Text for evaluating the robustness of text detectors in real complex scenes. Experimental results showed that the robustness of scene text detectors can be markedly improved by using our learnable scene transformation technique.

References

Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE Conference on CVPR*, pages 9365–9374, 2019.

Chris H Bahnsen, David Vázquez, Antonio M López, and Thomas B Moeslund. Learning to remove rain in traffic surveillance by using synthetic data. In *Proceedings of the 14th International Conference on Computer Vision Theory and Applications (visigrapp 2019)*, 2018.

G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE conference on CVPR*, pages 113–123, 2019.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Boris Epshtein, Eyal Ofek, and Yonatan Wexler. Detecting text in natural scenes with stroke width transform. In *2010 IEEE Computer Society Conference on CVPR*, pages 2963–2970. IEEE, 2010.

Alain Fournier, Don Fussell, and Loren Carpenter. Computer rendering of stochastic models. *Communications of the ACM*, 25(6):371–384, 1982.

Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems*, pages 7538–7550, 2018.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Daniel Ho, Eric Liang, Ion Stoica, Pieter Abbeel, and Xi Chen. Population based augmentation: Efficient learning of augmentation policy schedules. *arXiv preprint arXiv:1905.05393*, 2019.

LLC ImageMagick Studio. Imagemagick, 2008.

Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160. IEEE, 2015.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE Conference on CVPR*, pages 9404–9413, 2019.

Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In *Thirty-First AAAI*, 2017.

Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. *arXiv preprint arXiv:1911.08947*, 2019.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. In *Advances in Neural Information Processing Systems*, pages 6662–6672, 2019.

Pengyuan Lyu, Cong Yao, Wenhao Wu, Shuicheng Yan, and Xiang Bai. Multi-oriented scene text detection via corner localization and region segmentation. In *Proceedings of the IEEE Conference on CVPR*, pages 7553–7563, 2018.

Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 20(11):3111–3122, 2018.

Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

Jashojit Mukherjee, K Praveen, and Venugopala Madumbu. Visual quality enhancement of images under adverse weather conditions. In *2018 21st International Conference on ITSC*, pages 3059–3066. IEEE, 2018.

Lukas Neumann and Jiri Matas. A method for text localization and recognition in real-world images. In *ACCV*, pages 770–783. Springer, 2010.

Andrei Polzounov, Artsiom Ablavatski, Sergio Escalera, Shijian Lu, and Jianfei Cai. Wordfence: Text detection in natural images with border awareness. In *2017 IEEE ICIP*, pages 1222–1226. IEEE, 2017.

Dogancan Temel and Ghassan AlRegib. Traffic signs in the wild: Highlights from the ieee video and image processing cup 2017 student competition [sp competitions]. *arXiv preprint arXiv:1810.06169*, 2018.

Shangxuan Tian, Yifeng Pan, Chang Huang, Shijian Lu, Kai Yu, and Chew Lim Tan. Text flow: A unified text detection system in natural scene images. In *Proceedings of the IEEE ICCV*, pages 4651–4659, 2015.

Shangxuan Tian, Shijian Lu, and Chongshou Li. Wetext: Scene text detection under weak supervision. In *Proceedings of the IEEE ICCV*, pages 1492–1500, 2017.

Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

Igor Vasiljevic, Ayan Chakrabarti, and Gregory Shakhnarovich. Examining the impact of blur on recognition by convolutional networks. *arXiv preprint arXiv:1611.05760*, 2016.

Wenhai Wang, Enze Xie, Xiaoge Song, Yuhang Zang, Wenjia Wang, Tong Lu, Gang Yu, and Chunhua Shen. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In *Proceedings of the IEEE ICCV*, pages 8440–8449, 2019.

Yue Wu and Prem Natarajan. Self-organized text detection with minimal post-processing via border learning. In *Proceedings of the IEEE ICCV*, pages 5000–5009, 2017.

Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE Conference on CVPR*, pages 8818–8826, 2019.

Yongchao Xu, Yukang Wang, Wei Zhou, Yongpan Wang, Zhibo Yang, and Xiang Bai. Textfield: Learning a deep direction field for irregular scene text detection. *IEEE Transactions on Image Processing*, 2019.

Chuhui Xue, Shijian Lu, and Fangneng Zhan. Accurate scene text detection through border semantics awareness and bootstrapping. In *Proceedings of the ECCV*, pages 355–372, 2018.

Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Detecting texts of arbitrary orientations in natural images. In *2012 IEEE Conference on CVPR*, pages 1083–1090. IEEE, 2012.

Cong Yao, Xiang Bai, Nong Sang, Xinyu Zhou, Shuchang Zhou, and Zhimin Cao. Scene text detection via holistic, multi-channel prediction. *arXiv preprint arXiv:1606.09002*, 2016.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE ICCV*, pages 6023–6032, 2019.

He Zhang and Vishal M Patel. Densely connected pyramid dehazing network. In *Proceedings of the IEEE conference on CVPR*, pages 3194–3203, 2018.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zheng Zhang, Chengquan Zhang, Wei Shen, Cong Yao, Wenyu Liu, and Xiang Bai. Multi-oriented text detection with fully convolutional networks. In *Proceedings of the IEEE Conference on CVPR*, pages 4159–4167, 2016.

Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of the ieee conference on CVPR*, pages 4480–4488, 2016.

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on CVPR*, pages 5551–5560, 2017.