SPIN: Structure-Preserving Inner Offset Network for Scene Text Recognition

Chengwei Zhang[1*], Yunlu Xu[2*], Zhanzhan Cheng[3,2†],
Shiliang Pu[2‡], Yi Niu[2], Fei Wu[3] and Futai Zou[1‡]

[1]Shanghai Jiaotong University, China;   [2]Hikvision Research Institute, China;   [3]Zhejiang University, China
{cwzhang, zoufutai}@sjtu.edu.cn; {xuyunlu, chengzhanzhan, pushiliang, niuyi}@hikvision.com; wufei@cs.zju.edu.cn

Abstract

Arbitrary text appearance poses a great challenge in scene text recognition tasks. Existing works mostly handle with the problem in consideration of the shape distortion, including perspective distortions, line curvature or other style variations. *Rectification (i.e., spatial* transformers) as the preprocessing stage is one popular approach and extensively studied. However, *chromatic* difficulties in complex scenes have not been paid much attention on. In this work, we introduce a new learnable geometric-unrelated rectification, Structure-Preserving Inner Offset Network (SPIN), which allows the color manipulation of source data within the network. This differentiable module can be inserted before any recognition architecture to ease the downstream tasks, giving neural networks the ability to actively transform input intensity rather than only the spatial rectification. It can also serve as a complementary module to known spatial transformations and work in both independent and collaborative ways with them. Extensive experiments show the proposed transformation outperforms existing rectification networks and has comparable performance among the state-of-the-arts.

Introduction

Optical Character Recognition (OCR) has long been an important task in many practical applications. Recently, recognizing text in natural scene images, referred to as Scene Text Recognition (STR), has attracted great research interests due to the diverse text appearances and even the extreme conditions. Modern techniques in deep neural networks have been widely introduced (Shi, Bai, and Yao 2016; Shi et al. 2016; Lee and Osindero 2016; Shi et al. 2019; Wang and Hu 2017; Cheng et al. 2017, 2018; Liu et al. 2018a; Zhang et al. 2019; Yu et al. 2020).

Since existing methods are quite powerful for regular text recognition, reading irregular shaped text has become a challenging yet hot research topic for computer vision community. As spatial transformer network (STN) (Jaderberg et al.

---

*Authors contribute equally. Zhang did this work when he was an intern in Hikvision Research Institute.

†This work is completed under the supervision of Zhanzhan Cheng (contact email: chengzhanzhan@hikvision.com).

‡Corresponding author.

(a) Regular Scene Text

(b) Irregular Scene Text: Geometric Distortion

(c) Irregular Scene Text: Chromatic Distortion

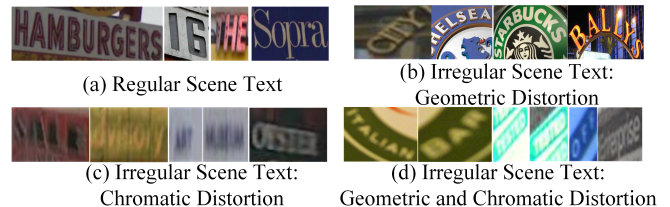(d) Irregular Scene Text: Geometric and Chromatic Distortion

Figure 1: Examples of regular and irregular scene text, where the problems of irregular case can be categorized into geometric distortion and chromatic distortion.

2015) was proposed, RARE (Shi et al. 2016) integrated it as a mainstreaming preprocessing procedure in the STR framework. Notice that the *rectification module* does not require any extra annotations yet is effective in an end-to-end network. Therefore, it was soon popularly utilized and further explored in later works (Bhunia et al. 2019; Baek et al. 2019; Shi et al. 2019; Zhan and Lu 2019; Luo, Jin, and Sun 2019; Yang et al. 2019). Although almost all the existing transformations are limited to geometric rectification, shape distortions do not account for all difficult cases in STR. Severe conditions resulting from intensity variations, poor brightness, shadow, background and imaging noise, even some insensible to human beings also contribute to hard samples for deep networks. In particular, corresponding to geometric distortion, we term these irregularities as *chromatic distortions*, which are very common and intractable but attracted rare attention in STR. Early work (Landau, Smith, and Jones 1988) stated that humans prefer to categorize objects based on their shapes, even the geometric distorted texts in Fig. 1(b) can be easily recognized on condition that *the shapes of objects are well preserved*. However, chromatic distortion leads to unclear shapes like in Fig.1(c) and (d), thus, recognition becomes much harder. To reduce the burden of STR, we are inspired to rectify the chromatic distortion to restore shapes of texts. That is, to design intensity manipulations towards variable distorted conditions.

Usually, chromatic distortions can be grouped into two cases (in Fig. 2) named *inter-pattern* and *intra-pattern* problems (explained later). Here we denote all the pixels with the same intensity as a *structure-pattern* (short in *pattern*). Assuming an ideal condition that the images to be identified are

(a) Rectifying
Inter-Pattern Problem

(b) Rectifying
Intra-Pattern Problem

Figure 2: Two types of chromatic difficulties. *Pattern* is defined as all the pixels of the same intensity in an image.

clearly composed of two distinct *pattern*s: the intended text *pattern* and the uncaring background *pattern*, networks can focus on the intrinsic shape or features of the text *pattern*. However, real scenes are far more complex than the above assumption, ubiquitous images are facing diverse chromatic challenges. In the detail of two chromatic distortion types, (a) *inter-pattern problem* means noise *pattern*s are close to the text *pattern*s (*e.g.,* poor contrast or brightness) or intensities of close text *pattern*s are dispersed.

Thus, it needs separate the text *patterns* and background *patterns*, and meanwhile aggregate characters into the uniform text *patterns*. As in Fig.2 (a), text becomes easier to recognize after intensities of 'L','O','V','E' and background are pulled away, while the intensities of the characters keep close. (b) *Intra-pattern problem* means text *pattern*s are interfered by the noise like shade, occlusion *etc*. As in Fig.2 (b), the shade on left-bottom is mixed with 'L', which should be obviated. Therefore, a good chromatic rectifier should have the ability to handle with both.

In this work, we propose a new chromatic rectifier called **SPIN**, which stands for *Structure-Preserving Inner Offset Network*. Here, we borrow the concept of *offset* from (Luo, Jin, and Sun 2019) intended for pixel-level spatial drift at birth, but we enrich it to a broader prospect, containing channel intensity offset (denoted in **inner offset**) and spatial offset (denoted in **outer offset**), respectively. The *inner offset* is designed to mitigate chromatic distortions, such as the mentioned *inter-* or *intra-pattern* difficulties, while the *outer offset* incorporates the geometric rectifications such as (Shi et al. 2016, 2019; Luo, Jin, and Sun 2019; Zhan and Lu 2019; Yang et al. 2019). Note that it is the first to focus on the *inner offset* of scene text, as the proposed SPIN mitigates chromatic distortions for a clearer shape of object. Specifically, it consists of two components: the *Structure Preserving Network (SPN)* and the *Auxiliary Inner-offset Network (AIN)*. SPN is responsible for alleviating the irregularities caused by *inter-pattern* problems. And AIN is an accessory network to distinguish the irregularities aroused by *intra-pattern* problems. These two components can complement each other. The rectified images will thus become visually clearer for deep networks in intrinsic shape of texts and then easier to recognize. Moreover, we explore the integration of both the inner and outer offsets by absorbing geometric rectification (*i.e.,* TPS (Shi et al. 2019; Baek et al. 2019)) into SPIN. Experiments verify that, the unified inner and outer offsets in a single transformation module will assist each other leading to even better recognition. The contributions can be summarized as:

(1) To the best of our knowledge, it is the first work to handle with the *chromatic distortions* in STR tasks, rather than the extensively discussed spatial ones. We also introduce the novel concept of *inner* and *outer* offsets in rectification networks and propose a novel SPIN to rectify the images with chromatic transformation.

(2) The proposed SPIN can be easily integrated into neural networks and trained in an end-to-end way without additional annotations and extra losses.

Unlike STN relying on tedious initialization schemes (Shi et al. 2016; Zhan and Lu 2019) or stage-wise training procedures (Yang et al. 2019), SPIN requires **no** need of sophisticated initialization or any network tricks, enabling it to be a more flexible module.

(3) The proposed SPIN achieves impressive effectiveness in multiple STR tasks. And the combination of chromatic and geometric transformations has been experimentally proved to be practicable, which further outperforms existing techniques by a large margin.

Related Work

Scene Text Recognition. Text recognition in natural scene images is one of the most important challenges in computer vision and many methods have been proposed. Conventional methods were based on hand-crafted features including sliding window methods (Wang, Babenko, and Belongie 2011; Wang and Belongie 2010), connected components (Neumann and Matas 2012), histogram of oriented gradients descriptors (Su and Lu 2014), and strokelet generation (Yao et al. 2016) *etc*. In recent years, deep neural network was dominating the area. As recurrent neural networks (RNNs) (Hochreiter and Schmidhuber 1997; Chung et al. 2014) were introduced and combined with CNN-based methods(Shi, Bai, and Yao 2016), more sequence-to-sequence models (Lee and Osindero 2016; Cheng et al. 2017; Wan et al. 2019; Lyu et al. 2019; Wang et al. 2020) were proposed. The attention mechanism was applied to a stacked RNN on top of the recursive CNN (Lee and Osindero 2016) and some revisions on the attention (Cheng et al. 2017) appeared. Recently, works further advanced the encoder-decoder structure with 2-Dimensional(2-D) ones including 2-D CTC (Wan et al. 2019), 2-D attention (Li et al. 2019; Lyu et al. 2019) and character-based segmentation before semantic learning (Wang et al. 2020; Yu et al. 2020). As modern techniques are powerful to deal with regular texts in scenes, irregular text recognition is still posing a challenging task. Existing methods include decoder on 2-D featuremaps (Li et al. 2019; Wan et al. 2019; Wang et al. 2020; Yu et al. 2020), character-level supervision and guidance (Yang et al. 2017), generator-based network (Liu et al. 2018b), and rectifications (Shi et al. 2016, 2019; Luo, Jin, and Sun 2019; Zhan and Lu 2019; Yang et al. 2019).

Rectification Network. Among above-mentioned approaches handling with *irregularity* of scene texts, rectification-based approaches are popularly equipped before recognition (Shi et al. 2016, 2019; Luo, Jin, and Sun 2019; Zhan and Lu 2019; Yang et al. 2019). A distinct advantage is that they do not require any additional annotations or loss functions, which is flexible yet effective (Baek et al. 2019; Chen et al. 2020). In this area, all the existing rectifications are limited to the geometric irregularity. Following

3306

Spatial Transformer Network (STN) (Jaderberg et al. 2015), RARE (Shi et al. 2016, 2019) first used thin-plate-spline (Bookstein 1989) transformation and regressed the fiducial transformation points on curved text. Similarly, ESIR (Zhan and Lu 2019) iteratively utilized a new line-fitting transformation to estimate the pose of text lines in scenes, and ScRN (Yang et al. 2019) improved the control point prediction in STN. MORAN (Luo, Jin, and Sun 2019) rectified multi-object using predicted offset maps. Different from the image-level rectification, STAR (Liu et al. 2016) proposed a character-aware network by using an affine transformation network to rotate each detected characters into regular ones. Here, we also work on designing rectification modules to ease downstream recognition. Differently from all the existing works, we point out that the *chromatic rectification* is also important. Our proposed network aims to rectify the input patterns from a broader perspective, and also handle with the color or brightness difficulties in STR tasks.

Methods

Chromatic Transformation

The proposed chromatic transformation *SPIN* consists of two components: the Structure Preserving Network (SPN) and the Auxiliary Inner-offset Network (AIN).

Structure Preserving Network (SPN) Inspired by Structure Preserving Transformation (SPT) (Peng, Zheng, and Zhang 2019), which cheats modern classifiers by degenerating the visual quality of chromaticity, we find that this kind of distortion is usually fatal to the deep-learning-based classifiers. The potential ability of controlling the chromaticity enlightens us to shed light on an even broader application conditioned on proper adaptation. In particular, we find that SPT-based transformations could also rectify the images of color distortions by intensity manipulation, especially for examples like Fig. 1 (c) and (d). Formally, given the input image $x \in \mathcal{I}$, let $x' \in \mathcal{I}'$ denote the transformed image. A transformation $\mathcal{T}$ on $\mathcal{I}$ is defined as follows:

$$x'(i, j) = \mathcal{T}[x(i, j)], \qquad (1)$$

where $x(i, j)$ or $x'(i, j)$ is the intensity of the input or output image at the coordinate (i, j). Specifically, the general form of SPT is defined to be a linear combination of multiple power functions:

$$x' = \mathcal{T}(x) = sigmoid(\sum_i \omega_i x^{\beta_i}), \qquad (2)$$

where β_i is the exponent of the i-th basis power function, ω_i is the corresponding weight. With normalized the intensity of each pixel in [0,1], total intensity space can be scattered into K subspace. Each space can be modeled by an exponential function under certain linear constraint as $y_i = x_i^{\beta_i}$, where $x_i = \frac{i}{2(K+1)}$, $y_i = 1 - \frac{i}{2(K+1)}$, i=1,2,3,...,K+1. Then 2K+1 parameters in pair can be formulated as:

$$\beta_i = \begin{cases} round(\dfrac{log(1 - \frac{i}{2(K+1)})}{log\frac{i}{2(K+1)}}, 2), 1 \le i \le K+1 \\ round(\dfrac{1}{\beta_{i-(K+1)}}, 2), K+1 < i \le 2K+1. \end{cases} \qquad (3)$$

These β exponents can be chosen based on domain or fixed in advance for simplicity. K defines the complexity of transformation intensities. A larger K will support a more complex and fine-grained chromatic space. ω is generated from the input image with convolutional blocks. Therefore, the transformation based on SPN can be summarized as:

$$x' = \mathcal{T}(x) = sigmoid(\sum_i W^s_{(i)} \mathcal{H}_{\mathcal{SPN}}(x) x^{\beta_i}), \qquad (4)$$

where W^s and $\mathcal{H}_{SPN}(\cdot)$ are part of Block8 weights and the feature extractor, respectively. $W^s \mathcal{H}_{SPN}(\cdot)$ is a *(2K+1)* dimensional segment from Block8 output of totally *(2K+2)* dimensions, as shown in Fig 3(a).

Layers	Configurations		Output	
	SPN	AIN	SPN	AIN
Block1	Conv(32, 3, 2, 2)		50×16	
Block2	Conv(64, 3, 2, 2)		25×8	
Block3	Conv(128, 3, 2, 2)		12×4	
Block4	Conv(256, 3, 2, 2)	Conv(16, 3, 2, 1)	6×2	6×2
Block5	Conv(256, 3, 2, 2)	Conv(1, 3, 1, 1)	3×1	6×2
Block6	Conv(512, 3, 1, 1)	-	3×1	-
Block7	Linear(256)	-	256	-
Block8	Linear(2K+2)	-	2K+2	-

Table 1: The architecture of SPN and AIN. The parameters in Conv($\cdot$) are filter numbers, kernel, stride and padding size of convolution orderly. For linear blocks, the value in Linear($\cdot$) means output channel numbers. Weights from Block1 to Block3 are sharing.

In essence, **Structure-Preserving** is realized by filtering the intensity level of input images. All the pixels with the same intensity level in the original image have the same intensity level in the transformed image, where the set $\{(i, j)|x(i, j) = c\}$ with intensity level c is defined as *structure-pattern*s. Intuitively, we propose the SPN for rectifying chromatic distortions by taking advantage of this singleton-based pixel-wise transformation in *two aspects*: (1) Separating useful structure-patterns from the harmful ones by injecting them to different intensity levels, which will be likely to generate better contrast and brightness. (2) Aggregating different levels of structure-patterns by mapping them to close intensity levels, which will be beneficial to alleviate fragments, rendering a more unified image. These are suitable for handling with *inter-pattern* problems (as Fig.2 (a)) but powerless against the other.

Auxiliary Inner-offset Network (AIN) Since SPN tries to separate and aggregate the specific *structure-pattern*s by exploiting the spatial invariance of words or characters, it inexplicably assumes that these patterns are under inconsistent intensities, namely different levels of *structure-pattern*s. However, it does not consider the disturbing patterns can have similar intensity with the useful ones, noted as *pattern confusion*, causing *intra-pattern* problems (as Fig.2 (b)).

To handle with the difficulty, we borrow the concept of *offset* from geometric transformation (Luo, Jin, and Sun 2019) while we decouple geometric and chromatic offsets for better understanding of the rectification procedures. SPN will generate **chromatic offsets**(namely *inner-offsets*) on
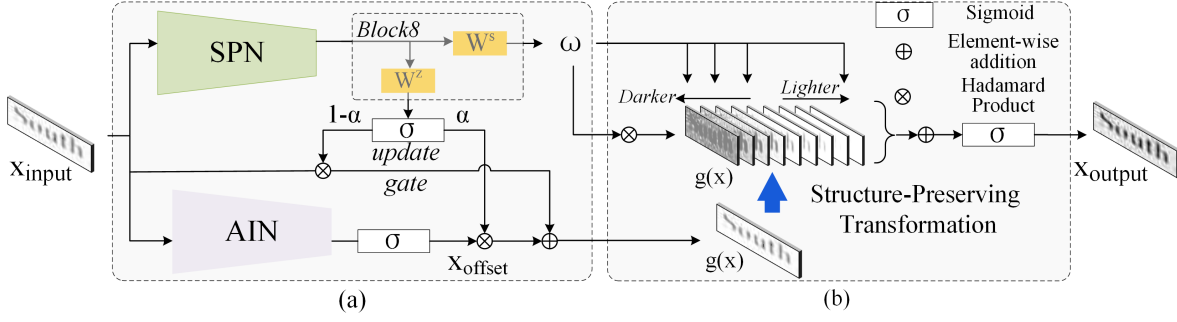
Figure 3: The overview of SPIN. (a) The input image x would first be feed into a well-designed network, and output the updated image x and a group of weights respectively. (b) A structure-preserving transformation is performed on the updated image with the generated weights.

each coordinate. As mentioned above, solely SPN is limited by the intrinsic *pattern confusion* problems. Here, we design an auxiliary inner-offset network (AIN) to assist with the consistent intensity for patterns. The *auxiliary inner-offset* is defined as:

$$g(x) = (1 - \alpha) \circ x + \alpha \circ x_{\text{offsets}}, \qquad (5)$$

where

$$\alpha = sigmoid(W^z \mathcal{H}_{\mathcal{SPN}}(x)), \qquad (6)$$

$$x_{\text{offsets}} = W^a \mathcal{H}_{\mathcal{AIN}}(x). \qquad (7)$$

W^z is the trainable scalar which is partial output from Block8 in Fig 3(a). Similarly, W^a is the trainable parameter in AIN module and $\mathcal{H}_{\mathcal{AIN}}(\cdot)$ is feature extractor sharing the first 3 blocks with SPN. $\circ$ is hadamard product. We design a learnable update gate α, which can receive signals from SPN and perceive the difficulty of different tasks. It is responsible for controlling the balance between the input image x and the predicted auxiliary inner-offsets. $g(x)$ (or x) is the updated (or input) image. Here, x_{offsets} is predicted by AIN, whose architecture is also given in Table 1. As chromatic transformation is pixel-to-pixel mapping on each coordinate and thus requires no need of spatial shift. The AIN first divides the image into small patches and then predicts offsets for each. All the offset values are activated by $sigmoid(\cdot)$ and mapped into the size of input image via common upsampling (*e.g.,* bilinear interpolation). The auxiliary inner-offsets can mitigate the *pattern confusion* through slight intensity-level perturbations $x_{\text{offsets}}(i, j)$ on each coordinate (i, j). Assisted by AIN, the enhanced transformation will be performed on the updated images, and the comprehensive transformation $\hat{\mathcal{T}}$ can be formulated upgrading Equa. (2) by

$$\hat{x}' = \hat{\mathcal{T}}(x) = sigmoid(\sum_i \omega_i (g(x))^{\beta_i}). \qquad (8)$$

The detailed network and the overall structure are also illustrated in Table 1 and Fig. 3.

Geometric-Absorbed Extension

Spatial transformations (Liu et al. 2016; Shi et al. 2016, 2019; Zhan and Lu 2019; Luo, Jin, and Sun 2019) rectify

the location shift of patterns by predicting the corresponding coordinates, which will generate geometric offsets (namely *outer-offset*s). And then they re-sample the whole images based on the points, which could be described as:

$$\tilde{x}'(i', j') = \mathcal{S}(x, f(i', j')) = \mathcal{S}(x, (i, j)), \qquad (9)$$

where $(i, j)(or(i', j'))$ is the origin coordinate (or the coordinate adjusted by outer-offsets), $\mathcal{S}(x, \cdot)$ here stands for the sampler, which generates the transformed images by interpolating the neighbor pixels in $x(i, j)$, and f represents different forms of transformation function, *e.g.,* Affine (Liu et al. 2016; Shi et al. 2016), Thin-Plate-Spline(TPS) (Bookstein 1989; Shi et al. 2016, 2019; Yang et al. 2019), Line-Fitting Transformation (Zhan and Lu 2019). Differently, in the proposed chromatic transformation generate *inner-offsets* on each coordinate as Equa. (1). Both are independent modules used to ease downstream stages, while they can be integrated as the unified transformation as

$$\ddot{x}'(i', j') = \mathcal{S}(\hat{x}', f(i', j'))$$
$$= \mathcal{S}(\hat{\mathcal{T}}(x), f(i', j')). \qquad (10)$$

We call the uniform of chromatic and geometric rectification as Geometric-Absorbed SPIN (short in *GA-SPIN*). Note that the learnable parameters in both chromatic transformation $\hat{\mathcal{T}}(\cdot)$ and geometric transformations $f(\cdot)$ are simultaneously predicted by the unified network configurations in Table 1. When absorbing spatial rectifications like TPS, the only difference for GA-SPIN compared to SPIN is the output channel numbers of Block8 are set to $2K + 2 + N$ instead of $2K + 2$, as N is number of parameters for TPS (*e.g.,* 40).

Recognition Network and Training

For direct comparison on rectification, we adopt the mainstreaming recognition network (Baek et al. 2019) and all its configuration and setups. Specifically, for contextual feature, we use typical STR backbones of 7-layer of VGG (Simonyan and Zisserman 2015) or 32-layer ResNet (He et al. 2016) with two layers of Bidirectional long short-term memory (BiLSTM) (Hochreiter and Schmidhuber 1997) each of which has 256 hidden units as (Shi, Bai, and Yao 2016;

Cheng et al. 2017). Following (Baek et al. 2019), the decoder adopts an attentional LSTM with 256 hidden units, 256 attention units and 69 output units (26 letters, 10 digits, 32 ASCII punctuation marks and 1 EOS symbol).

Experiments

Dataset

Models are trained only on 2 public synthetic datasets MJSynth (MJ) (Jaderberg et al. 2014) and SynthText (ST) (Gupta, Vedaldi, and Zisserman 2016) without any additional dataset or data augmentation. We evaluate on 4 regular and 3 irregular datasets as follows according to the difficulty and geometric layout of the text (Baek et al. 2019).

IIIT5K (IIIT) (Mishra, Alahari, and Jawahar 2012) contains scene texts and born-digital images. It consists of 3,000 images for evaluation.

SVT (Wang, Babenko, and Belongie 2011) contains 647 images for evaluation, collected from Google Street View, some of which are noisy, blurry, or of low-resolution.

ICDAR2003 (IC03) (Lucas et al. 2003) contains 860 images for evaluation which was used in the Robust Reading Competition in the International Conference on Document Analysis and Recognition (ICDAR) 2003.

ICDAR2013 (IC13) (Karatzas et al. 2013) contains 1015 for evaluation which was used in the ICDAR 2013.

ICDAR2015 (IC15) (Karatzas et al. 2015) contains 2,077 text image patches are cropped scene texts suffer from perspective and curvature distortions, which was used in the ICDAR 2015. Researchers have used two different versions for evaluation: 1,811 and 2,077 images. We use 1,811 images in our discussions and report both results when compared with other techniques for fair comparisons.

SVTP (Quy Phan et al. 2013) contains 645 images for evaluation. Many of the images contain perspective projections due to the prevalence of non-frontal viewpoints.

CUTE80 (CT) (Risnumawan et al. 2014) has 288 cropped images for evaluation. Many of these are curved text.

Implementation

All images are resized to 32×100 before entering the network. $K = 6$ is the default setting. The parameters are randomly initialized using He *et al.* 's method (He et al. 2015) if not specified. Models are trained with the AdaDelta (Zeiler 2012) optimizer for 5 epochs with batch size = 64. The learning rate is set to 1.0 initially and decayed to 0.1 and 0.01 at 4-th and 5-th epoch, respectively.

NOTE that the inconsistent choices of datasets, network and training configurations make fair comparison and module-wise assessment of recent works more difficult. We evaluate the proposed rectification module with SOTAs rigorously using the same datasets, recognition network and training setups as the unified benchmark (Baek et al. 2019). All experiments are only with word-level annotations. The network is trained from the scratch end-to-end without stage-wise procedures or other tricks.

Effects of Rectification Module

To go deep into individual factors in the proposed model, we cumulatively enable each configuration *one-by-one* on top of a solid baseline. All models are trained on the MJ dataset with backbone of 7-layer VGG. Following models are all the same by default. Table 2 lists the accuracy in each configuration.

Chromatic Transformation. The item (a) poses the baseline (Shi et al. 2016) without the rectification module. (b) and (c) add the rectification module and verify the advantage of the chromatic rectification. Specifically, (b) enables the proposed SPIN w/o AIN and improves (a) by an average of 0.9%. And (c) moves forward to enable AIN with SPN, composing SPIN. Though only two additional convolution layers appended to (b), (c) obtains an average of 0.4% consistent improvement and a total of 1.2% improvement compared with the baseline. The enhancement is mainly from SVT, IC15 and CUTE80, where lots of images with chromatic distortions are included. In addition, the first 3 rows in Fig. 5, mostly the *inter-pattern* distortions are transformed to clearer shape through SPIN-based method, and the last row, seemingly the *intra-pattern* problems(*i.e.,* shadow), also show the obvious effect between SPIN over w/o AIN.

Chromatic versus Geometric Transformation. Table 2(d) shows the baseline (a) equipped with STN (Baek et al. 2019), which rectifies images via *outer-offset*s and achieved superior performance among the state-of-the-arts. Interestingly, the performance of SPIN in (e) is close to that of the geometric one. The difference lies in that, the former mainly improves the SVT and CUTE80 benchmarks, where large proportion of chromatically distorted images are included, while the latter mainly improves the IC15 and SVTP benchmarks, where rotated, perspective-shifted and curved images dominate. And it verifies the color rectification is comparably important with shape transformation and the two types of rectifiers possess their respective strengths.

Combining Chromatic and Geometric Transformation. We further explore whether inner-offsets assist the existing outer-offsets (Shi et al. 2016) and how these two categories can jointly work better. In (e), simply placing the SPIN (c) before the TPS-based STN (d) (denoted as *SPIN+STN*) without additional modification brings an average of 0.8% improvement, compared with configuration (d). It indicates that these two transformations are likely to be two complementary strategies, where the inner part was usually ignored by the former researchers. GA-SPIN in (f) unifies SPIN and TPS-based STN in a single transformer (illustrated in Section 3.2). In detail, the inner and outer offsets are predicted by two single networks in (e) and the total parameters are about 4M. While for (f), the pipelines are optimized using the unified network and the total parameters are 2.31M. Though cutting parameters and computation cost compared to the straightforward pipeline structure (e), it improves the results by 0.3% on average and gains a large margin by 1.1% compared to solely geometric transformers (d). It means GA-SPIN is a simple yet effective integration.

Rectification	Transformation	Offsets		Benchmark							
		Outer	Inner	IIIT	SVT	IC03	IC13	IC15	SVTP	CT	Avg
None	(a) None[†] (Shi et al. 2016)	No	No	82.4	81.9	91.2	87.3	67.4	70.9	61.6	77.5
Chromatic	(b) SPIN w/o AIN	No	Yes	82.1	84.2	91.9	87.8	67.3	72.3	63.5	78.4
	(c) SPIN	No	Yes	82.3	84.1	91.6	88.7	68.1	72.3	64.6	78.8
Geometric	(d) STN[†] (Baek et al. 2019)	Yes	No	82.7	83.3	92.0	88.5	69.6	74.1	62.9	79.0
Both	(e) SPIN+STN	Yes	Yes	83.6	84.4	92.7	89.2	70.9	73.2	64.6	79.8
	(f) GA-SPIN	Yes	Yes	84.1	83.2	92.3	88.7	71.0	74.5	67.1	80.1

Table 2: Comparison between the baseline and the model combined with different rectifications. All the models are evaluated based on accuracy under several benchmarks. The 'Offset' stands for enabling the corresponding offset modules. '†' indicates the model is tested by us under a fair setting.

Discussion

Discussion on Sensitivity to Weight Initialization. Proper weight initialization is necessary for geometric transformations (Shi et al. 2016, 2019; Zhan and Lu 2019), since highly geometrically distorted images will ruin the training of the recognition network. On the contrary, chromatic transformations are less likely to severely distort the images for the intrinsic merits of preserving the *structure-pattern*s. Random initialization works well in Table 2 (b) and (c), indicating insensitivity to weight initialization of our modules.
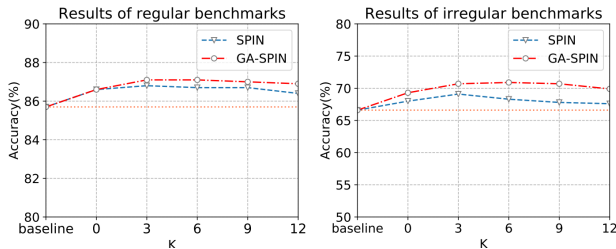

Figure 4: Results on different K. All the models are trained on the MJ dataset using VGG feature extractor. The line in coral, blue and red refer to the baseline without rectification module, with SPIN and with GA-SPIN, respectively. The results are evaluated with the mean accuracy.

Discussion on Different Values of K. We empirically analyze the performance under different K values, from 0 to

Transformation	Accuracy(%)		#Para.
	ResNet-MJ	ResNet	$(\times 10^6)$
None	69.0	74.9	0
STN (2019)[†]	71.2(+2.2)	76.9(+2.0)	1.68
ASTER (2019)	70.6(+1.6)*	78.0(+3.1)*	-
ESIR (2019)	72.7(+3.7)*	79.9(+5.0)*	-
ScRN (2019)	**73.0(+4.0)***	82.3(+7.4)*	2.62
SPIN	71.0(+2.0)	81.3(+6.4)	2.30
GA-SPIN	**73.0(+4.0)**	**84.5(+9.6)**	2.31

Table 3: Comparison with rectification methods. The values are mean accuracy of 3 irregular benchmarks. '-MJ' means only using MJ dataset. '*' denotes the results with deeper 45/50-layer than our 32-layer ResNet setting. '†' indicates the model is tested by us under the same setting for fairness.

Figure 5: Visualization of rectification with chromatic problems, unbalanced contrast (row 1-2), low brightness (row 3) and shadow (row 4),respectively. Samples here and in Fig. 6 are all from SVT, IC15 and CUTE80 test sets.
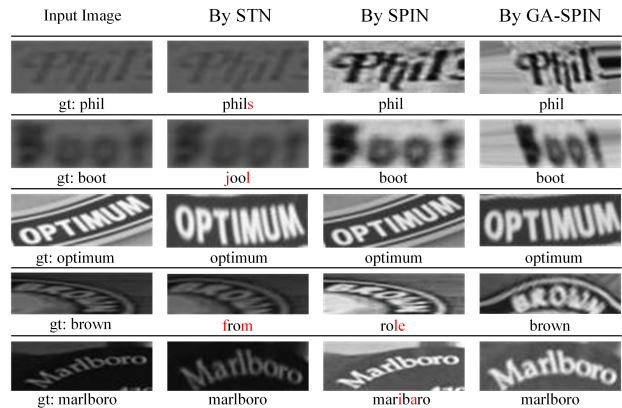

Figure 6: Examples of the rectification effects and prediction. "gt" is the ground truth. Predictions are shown under the images. Red are mistakenly recognized characters.

12, demonstrated in Fig. 4 with mean accuracy under regular and irregular cases. Both SPIN and GA-SPIN steadily improve the performance when K becomes larger to some extent. We do not observe additional performance gain when K is unreasonably large as the smaller K indicates simpler transformation and the larger K may result in complexity and redundancy. While K=3 seems enough for regular text, larger (*e.g.,* K=6,9) is better for irregular benchmarks.

Discussion on Traditional Image Preprocessing. As chromatic difficulties are rarely seen in STR networks, we consider traditional image processing techniques for col-

3310

Methods	Rec	ConvNet	Training data	Supervision	Benchmark							
					IIIT	SVT	IC03	IC13	IC15a	IC15b	SVTP	CT
Jaderberg (2016)		VGG	MJ	Word	-	80.7	93.1	90.8	-	-	-	-
Shi *et al.* (2016)		VGG	MJ	Word	78.2	80.8	89.4	86.7	-	-	-	-
Shi *et al.* (2016)	√	VGG	MJ	Word	81.9	81.9	90.1	88.6	-	-	71.8	59.2
Lee *et al.* (2016)		VGG	MJ	Word	78.4	80.7	88.7	90.0	-	-	-	-
Shi *et al.* (2019)	√	ResNet	MJ+ST	Word	93.4	89.5	94.5	91.8	76.1	-	78.5	79.5
Zhan *et al.* (2019)	√	ResNet	MJ+ST	Word	93.3	90.2	-	91.3	76.9	-	79.6	83.3
Baek *et al.* (2019)	√	ResNet	MJ+ST	Word	87.9	87.5	94.9	92.3	77.6	71.8	79.2	74.0
Luo *et al.* (2019)	√	ResNet	MJ+ST	Word	91.2	88.3	**95.0**	92.4	-	68.8	76.1	77.4
Xie *et al.* (2019)		ResNet	MJ+ST	Word	-	-	-	-	-	68.9	70.1	82.6
Wang *et al.* (2020)		ResNet	MJ+ST	Word	94.3	89.2	**95.0**	93.9	-	74.5	80.0	84.4
Luo *et al.* (2020)		ResNet	MJ+ST	Word	-	-	-	-	-	76.1	79.2	84.4
Yue *et al.* (2020)		ResNet	MJ+ST	Word	**95.3**	88.1	-	**94.8**	-	77.1	79.5	**90.3**
Zhang *et al.* (2020)		ResNet	MJ+ST	Word	94.7	**90.9**	93.3	94.2	81.8	-	81.7	-
Cheng *et al.* (2017)		ResNet	MJ+ST	Word+Char	87.4	85.9	94.2	93.3	70.6	-	-	-
Liao *et al.* (2019)		ResNet	MJ+ST	Word+Char	91.9	86.4	-	91.5	-	-	-	79.9
Yang *et al.* (2019)	√	ResNet	MJ+ST	Word+Char	94.4	88.9	95.0	93.9	-	78.7	80.8	87.5
Wan *et al.* (2020)		ResNet	MJ+ST	Word+Char	93.9	90.1	-	92.9	79.4	-	83.3	79.4
Li *et al.* (2019)		ResNet	MJ+ST+P	Word	91.5	84.5	-	91.0	-	69.2	76.4	83.3
Yu *et al.* (2020)		ResNet	MJ+ST+P	Word	94.8	91.5	-	95.5	82.7	-	85.1	87.8
Hu *et al.* (2020)		ResNet	MJ+ST+P	Word	*95.5*	*92.9*	*95.2*	94.3	82.5	-	*86.2*	92.3
Yue *et al.* (2020)		ResNet	MJ+ST+P	Word	95.4	89.3	-	94.1	-	79.2	82.9	*92.4*
SPIN	√	ResNet	MJ+ST	Word	94.7	87.6	93.4	91.5	79.1	76.0	79.7	85.1
GA-SPIN	√	ResNet	MJ+ST	Word	95.2	**90.9**	94.9	**94.8**	**82.8**	**79.5**	**83.2**	87.5

Table 4: Comparison with SOTA methods. 'P' indicates additionally using extra synthetic or real datasets apart from MJ and ST. 'Word' and 'Char' in 'Supervision' column means word-level and character-level annotations. For IC15, a and b indicate 1811 and 2077 examples, respectively. 'Rec' indicates methods focused on rectifications. For fair comparison, reported results using additional private data or annotations are not taken into account (Top result in italic). Top accuracy is shown in bold.

or correction (*e.g.,* equalization, binarization, morphological operations). We try these techniques as preprocessing stage in our recognition task similarly in train and inference phase. However, they do not work well (even not better than baseline). We attribute it to 2 main reasons: (1) They suffer from signal loss through image processing operations before flowing into the deep network. (2) They are manually designed for better image quality according to human intuition and visual feeling, not adapted to variable conditions.

Discussion on Promotion to Geometric Transformation. As mentioned, the union of the two can further boost the results (Table 2(e) or (f)). Table 7 further shows that GA-SPIN even has superior effects not only in chromatic but also in the geometric perspective over the single STN. We attribute it to the chromatic promotion for the geometric transformation, similar in the 4th row (the word 'brown') in Fig. 6. It means the SPIN eases the burden for recognition networks through chromatic rectification, and thus helps the training of spatial transformation in GA-SPIN.

Discussion on Training Procedures. Existing rectification modules are not easy to train as complained in (Shi et al. 2019; Luo, Jin, and Sun 2019; Yang et al. 2019). The proposed SPIN requires no need of sophisticated initialization, or stage-wise training strategies. Through the whole training procedures (visualized after 3 and 5 epochs in Fig. 8), we find that even after 3 epochs SPIN is almost completely trained, while the geometric rectification are still on the way to perform better. It shows that compared to geometric distortions, SPIN is much easier to learn and converge.
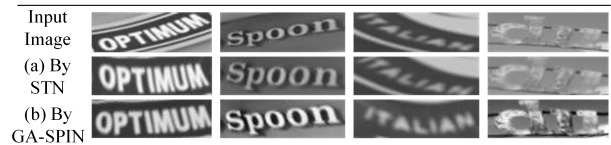

Figure 7: Examples of the rectification effects focused on the spatial distortions. Given input images with shape irregularities, (a) shows the results through STN-based transformation and (b) shows the results by the proposed GA-SPIN method. The downstream recognition networks after the rectification modules are the same for fair comparison.

Comparison with STN-based Methods

We compare SPIN and GA-SPIN with existing rectification networks in both quantitative and qualitative evaluation.

Quantitative Results. Table 3 lists the top results among existing geometric-based transformations (Shi et al. 2019; Zhan and Lu 2019; Yang et al. 2019) on irregular datasets with backbones based on ResNet and different training set (MJ/MJ+ST). Both SPIN and GA-SPIN achieve superior recognition performance compared to existing methods. Note the recent ScRN(Yang et al. 2019) also has impressive performance, but its backbone chooses FPN with ResNet-50 and uses additional explicit supervision with char-level annotations of SynText (Gupta, Vedaldi, and Zisserman 2016).

Qualitative Results. Some visualization in Fig. 5 clearly shows the proposed SPIN can rectify the chromatic distor-

		(a) GA-SPIN		(b) GA-SPIN
		Training on CUTE80		Training on ICDAR2015

Figure 8: Visualization of GA-SPIN training procedures. We show the input and the rectified images after 3 and 5 epoches.

tions by adjusting the colors. The typical chromatic inter-pattern (row 1-3) and intra-pattern (row 4) problems which geometric rectifications do not consider, can be mitigated.

Fig. 6 illustrates the rectification effects and text predictions by STN (Shi et al. 2019), SPIN and GA-SPIN, respectively. It shows rectifying chromatic distortion can improve the recognition results from two aspects: (1) It makes characters easier to recognize directly. The first 2 rows show that models tend to be misled by special symbols or similar shapes under severe chromatic distortions. At the same time, the proposed SPIN does not over-rectify or degrade images free of chromatic problems, such as the samples in the 3rd row. (2) Chromatic and geometric rectification can promote each other, as clearly shown in the last 2 rows of distorted images with both difficulties. Note in the 4th row, GA-SPIN has superior effects not only in chromatic but also in the geometric perspective over STN-based methods. We attribute it to the chromatic promotion for the geometric.

Comparison with the State-of-the-Art

Table 4 lists the reported the state-of-the-art results on STR tasks. The proposed SPIN-based rectification network achieves impressive performance across the 7 datasets.

Rectification-based Approaches. Networks focused on rectification are all denoted in '√' in Table 4. Although with only word-level annotations, SPIN-based already outperforms all the methods using spatial transformers (Shi et al. 2019; Baek et al. 2019; Zhan and Lu 2019; Yang et al. 2019) on overall performance(especially in more challenging IC15, SVTP and CT datasets), even when recent method (Yang et al. 2019) applied heavier backbone and additional char-level annotations. It points out that despite the widely focused geometric irregularity, color distortions are vital in these complex scenes and SPIN-based methods can effectively boost current performance.

Overall Performance among SOTAs. Note that even with strong settings like character-level annotations (Cheng et al. 2017; Liao et al. 2019) or deeper convolutional structures (*e.g.,* 45-layer ResNet (Shi et al. 2019; Luo, Jin, and Sun 2019)) and newly designed advanced framework (Wang et al. 2020), even compared to reported results using additional real or private data (*e.g.,* (Wan et al. 2020; Yang et al. 2019)), SPIN obtains satisfying performance and GA-SPIN shows significant promotion. Our methods only slightly fall behind (on solely IC03 dataset) the best reported (Wang
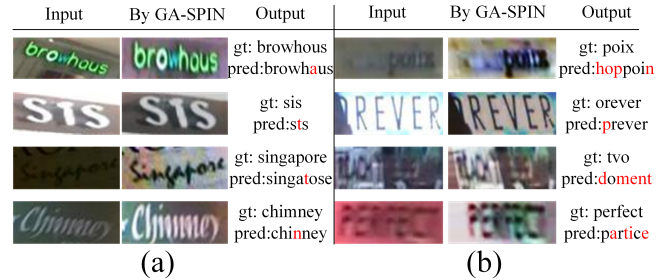

Input	By GA-SPIN	Output	Input	By GA-SPIN	Output
		gt: browhous pred:browhaus			gt: poix pred:hoppoin
		gt: sis pred:sts			gt: orever pred:prever
		gt: singapore pred:singatose			gt: tvo pred:doment
		gt: chimney pred:chinney			gt: perfect pred:partice
	(a)			(b)	

Figure 9: Some bad cases produced by our recognition system based on GA-SPIN. "gt" is the ground truth. "pred" is the prediction. Red are mistakenly recognized characters. (a) is from regular datasets. (b) is from irregular datasets.

et al. 2020; Luo, Jin, and Sun 2019) by 0.1%, and we attribute the reason to stronger feature extractor (*i.e.,* deeper convolution and multi-scale features) and advanced decoupled attention mechanism (Wang et al. 2020). Advanced networks bring strong ability on regular recognition but fall behind GA-SPIN by a large margin on more complex scenes (*e.g.,* IC15, SVTP and CUTE80). In overall performance on 7 benchmarks, our proposed GA-SPIN obviously surpasses all the existing methods under fair comparison.

Failure Cases. Failure cases are shown in Fig. 9. Similar characters like 'o' and 'a', 'i' and 't' in the first two images are still challenging in recognition. More difficulties include but not limit to shield (*e.g.,* "orever" in Fig. 9 (b)), stain (*e.g.,* "poix"), blur (*e.g.,* "perfect") and severe complex background (*e.g.,* "tvo") in real world. These cannot be directly solved by such general yet flexible transformations like STN-based or our SPIN-based networks, left unsolved.

Conclusion

This paper proposes a novel idea of chromatic rectification in scene text recognition. The proposed SPIN allows the channel intensity manipulation of data within the network, giving neural networks the ability to actively transform the input color for clearer patterns. Extensive experiments show its benefits the performance, especially in more complex scenes. It is also verified that geometric and chromatic rectifications can be unified into GA-SPIN rather than pipeline of two modules, which further boosts the results and outperform the existing methods by a large margin.

References

Baek, J.; Kim, G.; Lee, J.; Park, S.; Han, D.; Yun, S.; Oh, S. J.; and Lee, H. 2019. What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis. In *ICCV*, 4714–4722.

Bhunia, A. K.; Das, A.; Bhunia, A. K.; Kishore, P. S. R.; and Roy, P. P. 2019. Handwriting Recognition in Low-Resource Scripts Using Adversarial Learning. In *CVPR*, 4767–4776.

Bookstein, F. L. 1989. Principal warps: Thin-plate splines and the decomposition of deformations. *TPAMI* 11(6): 567–585.

Chen, X.; Jin, L.; Zhu, Y.; Luo, C.; and Wang, T. 2020. Text Recognition in the Wild: A Survey. *CoPR, abs/2005.03492* .

Cheng, Z.; Bai, F.; Xu, Y.; Zheng, G.; Pu, S.; and Zhou, S. 2017. Focusing attention: Towards accurate text recognition in natural images. In *ICCV*, 5076–5084.

Cheng, Z.; Xu, Y.; Bai, F.; Niu, Y.; Pu, S.; and Zhou, S. 2018. Aon: Towards arbitrarily-oriented text recognition. In *CVPR*, 5571–5579.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoPR, abs/1412.3555* .

Gupta, A.; Vedaldi, A.; and Zisserman, A. 2016. Synthetic data for text localisation in natural images. In *CVPR*, 2315–2324.

He, K.; Zhang, X.; Ren, S.; and Jian, S. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *ICCV*, 1026–1034.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9(8): 1735–1780.

Hu, W.; Cai, X.; Hou, J.; Yi, S.; and Lin, Z. 2020. GTC: Guided Training of CTC Towards Efficient and Accurate Scene Text Recognition. In *AAAI*, 11005–11012.

Jaderberg, M.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Synthetic data and artificial neural networks for natural scene text recognition. *CoPR, abs/1406.2227* .

Jaderberg, M.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2016. Reading Text in the Wild with Convolutional Neural Networks. *IJCV* 116(1): 1–20.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial Transformer Networks. In *NeurIPS*, 2017–2025.

Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V. R.; Lu, S.; et al. 2015. ICDAR 2015 competition on robust reading. In *ICDAR*, 1156–1160. IEEE.

Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; i Bigorda, L. G.; Mestre, S. R.; Mas, J.; Mota, D. F.; Almazan, J. A.; and De Las Heras, L. P. 2013. ICDAR 2013 robust reading competition. In *ICDAR*, 1484–1493. IEEE.

Landau, B.; Smith, L. B.; and Jones, S. S. 1988. The importance of shape in early lexical learning. *Cognitive development* 3(3): 299–321.

Lee, C.-Y.; and Osindero, S. 2016. Recursive recurrent nets with attention modeling for ocr in the wild. In *CVPR*, 2231–2239.

Li, H.; Wang, P.; Shen, C.; and Zhang, G. 2019. Show, Attend and Read: A Simple and Strong Baseline for Irregular Text Recognition. In *AAAI*, volume 33, 8610–8617.

Liao, M.; Zhang, J.; Wan, Z.; Xie, F.; Liang, J.; Lyu, P.; Yao, C.; and Bai, X. 2019. Scene Text Recognition from Two-Dimensional Perspective. In *AAAI*, 8714–8721.

Liu, W.; Chen, C.; Wong, K.; Su, Z.; and Han, J. 2016. STAR-Net: a SpaTial attention residue network for scene text recognition. In *BMVC*, 4168–4176.

Liu, Y.; Wang, Z.; Jin, H.; and Wassell, I. 2018a. Synthetically supervised feature learning for scene text recognition. In *ECCV*, 435–451.

Liu, Y.; Wang, Z.; Jin, H.; and Wassell, I. J. 2018b. Synthetically Supervised Feature Learning for Scene Text Recognition. In *ECCV*, 449–465.

Lucas, S. M.; Panaretos, A.; Sosa, L.; Tang, A.; Wong, S.; and Young, R. 2003. ICDAR 2003 robust reading competitions. In *ICDAR*, 682–687. Citeseer.

Luo, C.; Jin, L.; and Sun, Z. 2019. MORAN: A Multi-Object Rectified Attention Network for Scene Text Recognition. *Pattern Recognition* 109–118.

Luo, C.; Zhu, Y.; Jin, L.; and Wang, Y. 2020. Learn to Augment: Joint Data Augmentation and Network Optimization for Text Recognition. In *CVPR*, 13743–13752.

Lyu, P.; Yang, Z.; Leng, X.; Wu, X.; Li, R.; and Shen, X. 2019. 2D Attentional Irregular Scene Text Recognizer. *CoRR* abs/1906.05708.

Mishra, A.; Alahari, K.; and Jawahar, C. 2012. Scene Text Recognition using Higher Order Language Priors. In *BMVC*, 1–11.

Neumann, L.; and Matas, J. 2012. Real-time scene text localization and recognition. In *CVPR*, 3538–3545.

Peng, D.; Zheng, Z.; and Zhang, X. 2019. Structure-Preserving Transformation: Generating Diverse and Transferable Adversarial Examples. *CoRR* abs/1809.02786.

Quy Phan, T.; Shivakumara, P.; Tian, S.; and Lim Tan, C. 2013. Recognizing text with perspective distortion in natural scenes. In *ICCV*, 569–576.

Risnumawan, A.; Shivakumara, P.; Chan, C. S.; and Tan, C. L. 2014. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications* 41(18): 8027–8048.

Shi, B.; Bai, X.; and Yao, C. 2016. An End-to-end Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition. *TPAMI* 39(11): 2298–2304.

Shi, B.; Wang, X.; Lyu, P.; Yao, C.; and Bai, X. 2016. Robust Scene Text Recognition with Automatic Rectification. In *CVPR*, 4168–4176.

Shi, B.; Yang, M.; Wang, X.; Lyu, P.; Yao, C.; and Bai, X. 2019. ASTER: An Attentional Scene Text Recognizer with Flexible Rectification. *TPAMI* 41(9): 2035–2048.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 770–778.

Su, B.; and Lu, S. 2014. Accurate scene text recognition based on recurrent neural network. In *ACCV*, 35–48. Springer.

Wan, Z.; He, M.; Chen, H.; Bai, X.; and Yao, C. 2020. TextScanner: Reading Characters in Order for Robust Scene Text Recognition. In *AAAI*, 12120–12127.

Wan, Z.; Xie, F.; Liu, Y.; Bai, X.; and Yao, C. 2019. 2D-CTC for Scene Text Recognition. *CoRR* abs/1907.09705.

Wang, J.; and Hu, X. 2017. Gated Recurrent Convolution Neural Network for OCR. In *NeurIPS*, 335–344.

Wang, K.; Babenko, B.; and Belongie, S. 2011. End-to-end scene text recognition. In *ICCV*, 1457–1464.

Wang, K.; and Belongie, S. 2010. Word spotting in the wild. In *ECCV*, 591–604. Springer.

Wang, T.; Zhu, Y.; Jin, L.; Luo, C.; Chen, X.; Wu, Y.; Wang, Q.; and Cai, M. 2020. Decoupled Attention Network for Text Recognition. In *AAAI*, 12216–12224.

Xie, Z.; Huang, Y.; Zhu, Y.; Jin, L.; Liu, Y.; and Xie, L. 2019. Aggregation Cross-Entropy for Sequence Recognition. In *CVPR*, 6538–6547.

Yang, M.; Guan, Y.; Liao, M.; He, X.; Bian, K.; Bai, S.; Yao, C.; and Bai, X. 2019. Symmetry-Constrained Rectification Network for Scene Text Recognition. In *ICCV*, 9147–9156.

Yang, X.; He, D.; Zhou, Z.; Kifer, D.; and Giles, C. L. 2017. Learning to Read Irregular Text with Attention Mechanisms. In *IJCAI*, 3280–3286.

Yao, C.; Bai, X.; Shi, B.; and Liu, W. 2016. Strokelets: A Learned Multi-Scale Mid-Level Representation for Scene Text Recognition. In *IEEE Trans. Image Process*, 2789–2802.

Yu, D.; Li, X.; Zhang, C.; Han, J.; and Ding, E. 2020. Towards Accurate Scene Text Recognition with Semantic Reasoning Networks. In *CVPR*, 12110–12119.

Yue, X.; Kuang, Z.; Lin, C.; Sun, H.; and Zhang, W. 2020. RobustScanner: Dynamically Enhancing Positional Clues for Robust Text Recognition. In *ECCV*.

Zeiler, M. D. 2012. ADADELTA: an adaptive learning rate method. *CoPR, abs/1212.5701* .

Zhan, F.; and Lu, S. 2019. ESIR: End-To-End Scene Text Recognition via Iterative Image Rectification. In *CVPR*, 2059–2068.

Zhang, H.; Yao, Q.; Yang, M.; Xu, Y.; and Bai, X. 2020. Efficient Backbone Search for Scene Text Recognition. *CoRR* abs/2003.06567.

Zhang, Y.; Nie, S.; Liu, W.; Xu, X.; Zhang, D.; and Shen, H. T. 2019. Sequence-To-Sequence Domain Adaptation Network for Robust Text Image Recognition. In *CVPR*, 2740–2749.