This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3171108, IEEE Transactions on Multimedia

Dual Relation Network for Scene Text Recognition

Ming Li, Bin Fu, Han Chen, Junjun He and Yu Qiao, *Senior Member, IEEE*

***Abstract*—Local visual and long-range contextual features yield two complementary cues for human reading text in natural scene. Existing scene text recognition methods mainly extract local features at a low level and then model long-range dependencies at a high level, this sequential pipeline may be sub-optimal to construct complete and effective representation. Except for high-level features, long-range contextual relation is of importance in low-level features as well since it can help separate different characters based on the intervals between characters and thus enhance the character features. To address this issue, we develop a dual relation module to extract complementary features in a parallel manner for scene text recognition, which consists of a local visual branch and a long-range contextual branch. The local visual branch employs a topological-aware operation to model intra-character characteristic and extract discriminative features of different characters. Meanwhile, the long-range contextual branch utilizes a simple but effective strategy to incorporate inter-character relations into feature maps. Our dual relation module is a plug-and-play block which can be easily incorporated into modern deep architectures. Experimental results demonstrate that our methods achieved top performance on several standard benchmarks. Code and models will become publicly available in the future.**

***Index Terms*—Scene Text Recognition, Scene Optical Character Recognition, Deep Learning**

I. Introduction

SCENE text recognition (STR) has become an active research topic in the multimedia community, which is a pivotal visual recognition task for many promising downstream applications, such as automatic driving [1], [2] and travel translator [3], [4]. To read text instances correctly, two types of complementary information, local visual information and long-range contextual information, both play significant roles in this task. For the STR task, the local visual features mainly represent intra-character relations for each individual character while long-range contextual features provide inter-character correlations between different characters. Therefore, extracting local visual features and modeling long-range contextual relations in an effective manner is an important issue for STR task.

Existing STR methods usually extract local features at a low level and model contextual correlation at a high level, which is a sequential manner [5], [6]. Specifically, a Convolution Neural Network (CNN) based feature extraction module is firstly employed to collect local visual relations. Then a contextual layer, mostly LSTM [7] or GRU, is utilized on the top of CNN to explore long-range contextual relations [8], representing the sequential relations between each character. Although this sequential feature extraction pipeline has improved recognition performance on many standard benchmarks, it may be sub-optimal to construct complete and effective representation, since some critical contextual clues are missing in above pipeline. For the local visual relation, since the topological structures are similar even with large shape and scale variance for the same character, traditional convolutional network cannot effectively extract topological relations among intra-character features. Topological structure is the position relations of the characters' pixels, which defines characters into specific classes. Moreover, for the long-range contextual information, existing models only encode above relations in the high-level features, which will lose some important contextual clues. For example, as shown in Fig. 1, characters in the same text instance usually have similar appearances compared with the background, which can cause intervals between characters. The characters, background and intervals, appear continuously and alternatively from the beginning to the end, which provides the low-level contextual priors to separate neighbor characters and enhance the forwarding features.

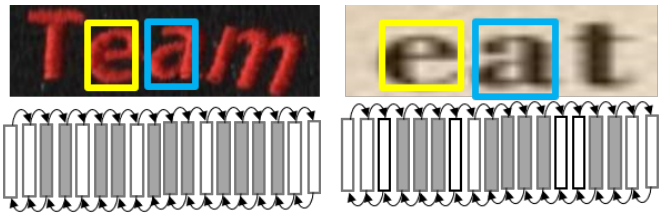


Fig. 1: Visualization of motivation. The first row is the text images, and characters in the same category are drawn in the rectangle of same color. The topological structure of same character is similar, which motivates us to build the topological-aware operator. The second row illustrates the motivation of long-range contextual branch, where the low-level contextual priors, such as characters and intervals, can be extracted to enhance the forwarding features. Each rectangle represents a vector and the arrows represent the flow of information. The shadow rectangles represent the positions of characters.

Based on the above observation, in order to effectively encode the low-level clues, we process long-range contextual and local visual relations simultaneously to construct complete representation. Therefore, we propose a dual relation module to extract local features and model contextual correlations in a parallel manner, consisting of a local visual branch

Manuscript received October 18, 2021; accepted April 19, 2022. This work is partially supported by the Joint Lab of CAS-HK, the Shenzhen Research Program (JSGG20191129141212311, RCJC20200714114557087), the Shanghai Committee of Science and Technology (Grant No. 21DZ1100100).

M. Li, B. Fu, H. Chen, J. He and Y. Qiao are with Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China. (email: {ming.li3@siat.ac.cn, bin.fu@siat.ac.cn, han.chen@siat.ac.cn, hejunjun@sjtu.edu.cn, yu.qiao@siat.ac.cn})

Y. Qiao is also with Shanghai AI Laboratory, Shanghai, China.

M. Li and B. Fu are equally-contributed authors.

Y. Qiao is the corresponding author.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3171108, IEEE Transactions on Multimedia

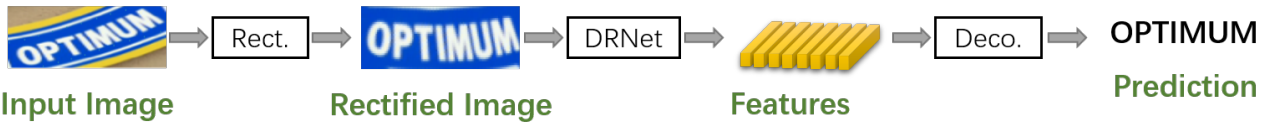


Fig. 2: Overall backbone. Rect., Fea. and Deco. represent Image Rectification, Feature Extraction and Final Decoding stage, respectively. In our current work, we propose Dual Relation Network (DRNet) to extract local feature and model long-range contextual relation in a parallel manner.

and a long-range contextual branch. Specifically, for the local visual branch, considering the visual characteristics of text symbols in STR task, a topological-aware operation is developed to extract discriminative intra-character relations. For the long-range contextual branch, instead of discarding vertical information directly, a simple but effective strategy is implemented to model inter-character contextual correlations. Vertical information is refer to the contextual information along vertical directions in the feature maps. Moreover, following the practice in modern network architecture [9], [10], we formulate our dual relation module into a residual block, termed as DR-block. Our DR-block is a plug-and-play module which can be utilized as a direct replacement of convolutional block in any neural network.

To verify the effectiveness of our proposed method, we implement DR-block on the classic scene text recognition platform [11] to build Dual Relation Network (DRNet) and conduct extensive experiments on various public benchmarks. Compared with our baseline, our DRNet improves recognition performance on all benchmarks with a large margin and obtains promising performance on a number of datasets. Moreover, we further replace the residual blocks in SAR [6] to build DRNet-SAR and obtain significant improvement, which demonstrates the generality of DR-block in two-dimensional attention based model.

The contributions of this paper are threefold :

- To include low-level cues in contextual information, we propose a dual relation block (DR-block) to extract local features and model long-range contextual correlations in a parallel manner.
- Our DR-block is a plug-and-play module and can be used as a direct replacement of convolutional block for any STR platforms to further boost recognition performance.
- We implement our DR-block on the classic scene text recognition platform, termed as DRNet, to verify the effectiveness of our proposed method. Our DRNet significantly improves recognition performance of the baseline model and achieves promising performance on a number of public datasets.

II. Related Work

In this section, we briefly review the recent progresses in scene text recognition and the self-attention mechanism. Scene text recognition model mostly follows an scene text detection model [12]–[14], which detects the text areas from a whole scene image, and recognizes text instances from the cropped text images.

A. Scene Text Recognition

Scene text recognition (STR) is an active research topic in multimedia community, which aims to translate the visual-form of information into other forms. Specifically, scene text detection models [15]–[18] locate the position of texts, and then the followed scene text recognition model translates the cropped text images into machine-readable symbols. Compared with the task of optical character recognition (OCR) [19], [20], scene text recognition tackles the texts appeared in the wild, which suffer from blur and distortion severely. Various methods have been developed to read text in the wild and in this section we will give a brief introduction of several state-of-the-art models, and a comprehensive study for the development of STR can be found in [21].

[22] divides modern STR methods into four common steps: The transformation stage first transforms text instances into the near-horizontal forms. Then a feature extraction stage, usually deep convolutional neural network, is employed to extract local visual information from the rectified images. A sequence modeling stage further extract long-range dependencies among different features. Finally, the prediction stage is utilized to generate prediction of text instances. Existing STR methods mainly focus on improving recognition performance on irregular text instances via the carefully-designed transformation stage or prediction stage. In the first case, [5], [11], [23]–[25] keep traditional one-dimensional attention-based model unchanged and employs a rectification module to convert irregular text instances to regular ones before recognition model. Inspired by Spatial Transformer Networks [26], ASTER [11] employs a Text Rectification Network to transform multi-orientation or curved text [27] into horizontal text and then employs sequence model to recognize text content in rectified instance. [25] employs several local attributes to generate more accurate text lines and then obtains control points by performing equidistantly sampling on them. LCSegNet [28] utilizes segmentation model to generate pixel-wise prediction for each character and employs conditional random field to smooth label assignments, which achieves promising performance in several public benchmarks. The second approach generalizes 1D prediction models into two-dimensional (2D) versions by developing 2D attention based decoder [29]–[31], where the sequence models employ complete geometric information to perform feature alignment. In addition, incorporating linguistic knowledge to refine recognition results has received much attention in recent studies. For example, [32] designs an iterative language model to correct recognition results, while [33] incorporates a dictionary as the language priors to refine the initial recognition results. Furthermore, exploring better

representation for characters is also a promising direction in STR task. [34] attempt to construct primitive representation on undirected graph and utilize GCN to obtain better representation for characters.

Different from above works, in this paper, we develop a robust feature extraction model, namely dual relation block, to extract complementary local and contextual features for STR task in a parallel manner. The dual relation block is a plug-and-play module and can be used as a direct replacement of convolutional block for any STR platforms to further boost recognition performance.

B. Convolution and Self-attention Operation

In our model, a Topological Aware operator is proposed for the local visual branch, instead of normal convolution, thus a brief introduction on self-attention operation is discussed below.

Convolution operation has been proposed for more than two decades, which is initially designed for recognizing handwritten digits [35]. With the impressive performance in image recognition task, it has become the dominated technique to extract visual features for various computer vision tasks. Since then, various extensions to the convolutional operation have been developed, which can be divided into two different directions. The first direction is to extract feature map in an efficient manner, such as group convolution [36] and depth-wise convolution [37], [38], while the second is to extract more powerful representations, such as dilated convolution [39] and deformable convolution [40].

Different from convolution operation in which the aggregation weights are fixed, self-attention operation adaptively aggregates visual features based on the pixel-wise relations. The self-attention model is designed for natural language processing initially, and has been widely employed in various computer vision tasks to extract image-level long-range interactions. For example, [41] employs self-attention to extract capturing long-range dependencies for object detection task while [42] further extends self-attention to model the semantic inter-dependencies in spatial and channel dimensions. Recently, several works [43], [44] pay attention to construct neural networks with self-attention operation by limiting the scope of self-attention to a local region, which have achieved promising results in image classification and segmentation tasks.

Since self-attention operation aggregates visual features according to the similarity relations, this operation can effectively preserve topological relations between different features, which is suitable to model intra-character relations in STR task. Based on this observation, we design a topological-aware operation to extract local visual information in this paper.

C. Contextual Feature Embedding

In recent years, contextual information has received much attention on various visual tasks, which can provide significant clues to distinguish different objects from their surroundings. Existing methods can be roughly divided into three categories based on the different feature extraction approaches. Firstly, several methods utilize the feature pyramid to extract multi-scale contextual information. For example, the DeepLab [45]–[47] adopt atrous spatial pyramid pooling (ASPP) module to enhance multi-scale contextual information by employing different dilated convolutions in a parallel fashion. Moreover, since the global average pooling operation will bring a significant enhancement for context information, PSPNet [48] develops a pyramid pooling module with different pooling sizes to collect useful global information. Secondly, the self-attention mechanism is utilized to enhance global context information by exploiting long-range dependencies between different pixels, such as DANet [42] and CCNet [49]. Finally, the gate mechanism is also employed to adaptively extract useful context information and filter the unrelated information under the carefully-designed operation. For example, Ding [50] proposes a context contrasted local feature which spotlights the local information in contrast to the context. ACNet [51] introduces a competitive fusion mechanism for global context and local context to capture the pixel-aware contexts information.

Unlike the above methods, for STR task, the long-range contextual information mainly exists in the direction of the text line. Moreover, due to the nature of language, the contextual connection are more important between each character. Therefore, our Long-range Contextual Branch compresses the feature map in vertical direction and then utilize the LSTM to extracts contextual information along the horizontal direction.

D. Compared with Related Works

In this subsection, we compare our current model with several similar works. Since the topological structures are similar for the same character, we propose the topological-aware (TA) operator to extract rotation-robust local features in our local visual branch.

There are several similar works pay attention to extracting the rotation-robust features. SIFT [52] is an classic handcrafted descriptor which extracts distinctive invariant features from images based on the spatial histogram of the image gradient. Though practical in the usage of image matching, it cannot extract rich and powerful visual information than deep network. The AON [29] encodes an arbitrarily-oriented text instance into the rotation-robust representation by weighted fusion the four feature sequences of four directions. In SLOAN [53], the rotation-aware representation is achieved by converting the rotation and scale of text instances from Cartesian coordinate space into the vertical and horizontal shift in log-polar space via the Dynamic log-Polar Transformer. Unlike the [29] and [53], our proposed TA operation can generate the strictly local rotation-invariant features for the predefined kernel regions, which is achieved by collecting local information based on the topological relation weights.

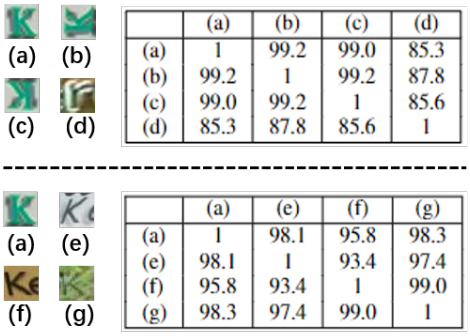
Moreover, several recent methods [43], [44], [54], [55] also attempt to replace traditional convolution blocks by the local self-attention blocks. For example, ViL [55] utilize sliding window based local self-attention to achieve linear complexity computational cost, which is similar with our TA operator. Although our TA operator can be regarded as the local self-attention and share the similar expression with some existing

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3171108, IEEE Transactions on Multimedia

works, the most important contribution of our current work is the parallel feature extraction pipeline, which simultaneously extracts local visual information and long-range contextual relations from two branches in a single block (DR block).

III. Method

In this section, we will give a detailed description on our model, the overall pipeline is shown in Fig. 2 which consists of three stages including Image Rectification, Feature Extraction and Final Decoding stage. Our current work mainly focuses on designing a dual relation module to construct complete and effective representation in Feature Extraction stage. This section is organized as follows: Firstly, the visual characteristics of STR task is analysed and the local visual branch is proposed to extract intra-character relations based on a topological-aware operation. Subsequently, a long-range contextual branch is proposed to model inter-character correlations. In section C, we formulate the local visual branch and long-range contextual branch as a Dual Relation Block, which can be regarded as a plug-and-play network structure for any STR platforms. Finally, we provide our overall pipeline to verify the effectiveness of our proposed model.



	(a)	(b)	(c)	(d)
(a)	1	99.2	99.0	85.3
(b)	99.2	1	99.2	87.8
(c)	99.0	99.2	1	85.6
(d)	85.3	87.8	85.6	1

	(a)	(e)	(f)	(g)
(a)	1	98.1	95.8	98.3
(e)	98.1	1	93.4	97.4
(f)	95.8	93.4	1	99.0
(g)	98.3	97.4	99.0	1

Fig. 3: Cosine similarities of different input images. In the upper part, (a) and (d) represent original single character images cut from real datasets, (b) and (c) are images rotated by 90 and 180 degrees from (a). In the lower part, (a)(e)(f)(g) represent different images of category "K". The tables on the image represent the cosine similarity between different input images.

A. Local Visual Branch

To begin with, we first analyse the visual characteristic in STR task. For same characters, the topological structures are similar even with large shape and scale variance as shown in Fig . 3(a)(e)(f)(g), while different characters contain various structures. For example, as shown in Fig . 3, after rotation (in Fig . 3(b) and (c)), topological structures are similar with the origin character (in Fig. 3(a)) and have a large difference with other characters (such as "r" in Fig. 3(d)). Therefore, to extract robust intra-character relations, the key problem is to make the features closed with each other for the symbols with similar topological structures (eg: different "K" in Fig. 3(a),(b),(c)), while keeping discriminable for different symbols.

The existing convolutional operation is problematic for modeling topological structures since the weights are fixed and cannot calculate features adaptively according to different structures. To handle this issue, we propose a topological-aware operation (TA operation) to adaptively extract intra-character relations in a local region. For each position i in input feature map $\mathbf{X}$, the topological structures will be perceived in the predefined 3×3 local regions $R(i)$ and topological relation weights $\alpha_{i,j}$ are calculated by the following equation:

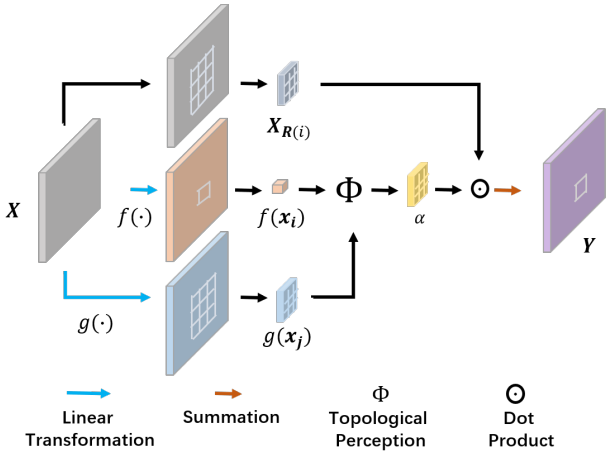


Fig. 4: Structure of Topological-Aware (TA) Operation. X and Y represent input and output features. $X_{R(i)}$ represents the input features in the local region $R(i)$. $g(\cdot)$ and $f(\cdot)$ represent two different linear transformations. α represents topological relation weights. Only the TA operation for y_i is shown in this figure.

$$\alpha_{i,j} = \Phi(x_i, x_j) = \frac{1}{n} \frac{\exp\left(f(x_i) \cdot g(x_j)\right)}{\langle \exp\left(f(x_i) \cdot g(x_j)\right) \rangle_j}, \tag{1}$$

where x_i and x_j are features vectors, $j \in R(i)$. Pixel i represents the center pixel of the operation kernel. $R(i)$ represents the region centered on pixel i (including i) and j is the pixel in $R(i)$. n is the number of pixels in the region $R(i)$. $f(\cdot)$ and $g(\cdot)$ are two different linear transformations. The $\langle \cdot \rangle_j$ denotes the average operator with respect to j in the predefined region $R(i)$. The topological relation weights encode the structure-specific relations between the center pixel i and the pixels in region $R(i)$. Therefore, for a specific position i, the topological-aware operation extracts local information according to the topological relation weights, which can be formulated as:

$$y_i = \sum_{j \in R(i)} \alpha_{i,j} \cdot x_j. \tag{2}$$

To illustrate the topological-aware characteristic of the proposed operation, we apply Eq. 2 to calculate the response map for images in Fig. 3 and compare the similarities among them. Specifically, we randomly selected some single character images from Syntext dataset, and then initialize one layer of TA-operator with a large kernel. Then we input the single-character image to the TA-operator and get the output feature, which can be used to calculate cosine similarities. As shown in the table, our topological-aware operation produces similar features for the same character and discriminative features for different characters. Therefore, we believe our local visual

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3171108, IEEE Transactions on Multimedia

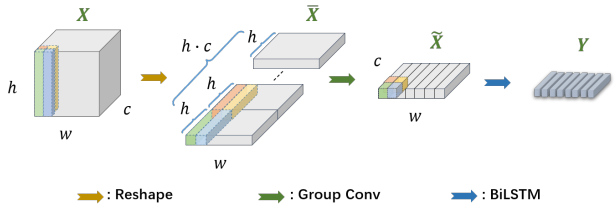


Fig. 5: Structure of group convolution (GC) version contextual branch. X and Y are input and output feature, $\bar{X}$ and $\tilde{X}$ are two intermediate features. h, w and c represent the height, width and channel numbers of the input feature.

branch can extract discriminable features for different characters in a more effective manner.

Finally, we employ the proposed topological-aware operation followed by a 1×1 convolutional layer as our local visual branch to extract intra-character features.

B. Long-range Contextual Branch

As discussed in previous section, contextual relation also plays a significant role in scene text recognition task. Existing STR methods process contextual relation in a sequential manner, which is a sub-optimal solution. In this section, we introduce a simple but effective contextual branch to model the long-range dependencies. Unlike the traditional contextual extractor [8], [11] which discards vertical information directly, in this paper, two different approaches, group convolution based (GC-based) and normal convolution based (NC-based), are proposed to adaptively compress vertical information, and then a BiLSTM layer is employed to extract inter-character relations. We first introduce feature compression approaches in the following.

In the group convolution based approach as shown in Fig. 5, we first reshape the input feature map $X \in R^{h \times w \times c}$ into the form $\bar{X} \in R^{1 \times w \times (h \times c)}$. Under this strategy, the position correspondence can be formulated as:

$$\bar{X}_{1,j,(k-1)\cdot h+i} = X_{i,j,k} \tag{3}$$

where i, j, k represent the index of height, width and channel respectively. From this operation, each features in $\bar{X}$ can be naturally separated into c groups with h channels. Therefore, we utilize a 1×3 group convolution to extract and compress vertical information into $\tilde{X} \in R^{1 \times w \times c}$. In normal convolution based (NC-based) approach, a $h \times 3$ convolutional operation is deployed to collect near-neighbor information along horizontal direction and obtain features into $\tilde{X} \in R^{1 \times w \times c}$. The reason we introduce group convolution is we need to compress the features of different scales to the height of 1 while the regular convolution will introduce more FLOPs and parameters and maxpooling will discard most vertical information.

Finally, a BiLSTM operation is utilized to model long-range relations along horizontal direction.

C. Dual Relation Block

Following the practice in modern deep network architecture [9], [10], we combine the proposed local visual and long-range contextual branches into an uniform network module, termed as DR-block. The structure of our DR-block is shown in Fig. 6. A 1×1 convolutional operation is firstly applied to project the input feature into a suitable feature space. Then the resulted features are passed to local visual and long-range contextual branches to extract intra-character relations and model inter-character correlations in a parallel manner. Afterwards they are combined via pixel-wise summation, followed by another 1×1 convolutional operation to better fuse the features. Finally, the input feature maps are added to the resulted feature maps via a residual connection [9].

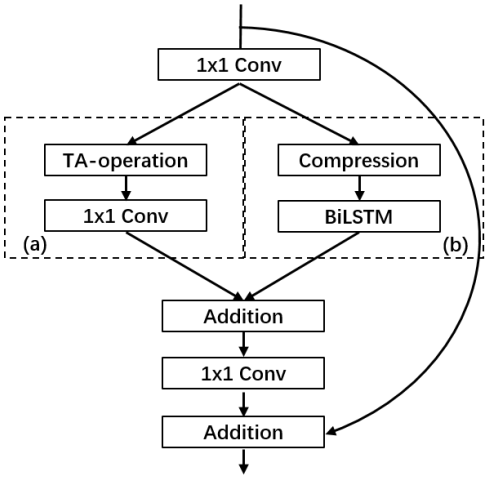


Fig. 6: Structure of DR-block. (a) is local visual branch and (b) is long-range contextual branch. Compression in (b) can be normal convolution (NC) version or group convolution (GC) version. Addition represents element wise addition.

D. Overall Pipeline

In this paper, to verify the effectiveness of our DR-block, we implement our plug-and-play module on classical scene text recognition platform ASTER [11] to build the Dual Relation Network (DRNet). For fairly comparison, we perform two modifications on ASTER: 1. Replacing the ResNet blocks [9] in feature extraction stage with our DR-block except for the first downsamping block. 2. Removing contextual stage [8], [22] since the contextual relations have been encoded in our DR-block. The overall recognition network configurations are shown in TABLE I.

After above modifications, the overall pipeline is shown in Fig. 2, which can be divided into three stages, including Image Rectification, Feature Extraction and Final Decoding stage. The image rectification stage is employed to rectify irregular and curved text into the horizontal ones, which directly regresses the control points for Thin Plate Spline (TPS) [56] without any supervision. For the feature extraction stage, we utilize our proposed DR-block to replace ResNet-block to extract intra- and inter-character relations simultaneously. For the character prediction stage, the RNN-based attentional decoder [11] is employed to predict each character based on the former prediction, which can be formulated as follows:

	Stages	Out Size	Configurations
Encoder	Stage 0	$32 \times 32 \times 100$	3×3 conv, s 1×1
	Stage 1	$32 \times 16 \times 50$	$[\ 1 \times 1$ conv, 3×3 conv $]\ \times 1$, s 2×2 $\begin{bmatrix} 3 \times 3 \text{ TA}, \\ 16 \times 3 \text{ conv}, \quad BiLSTM \end{bmatrix} \times 2$
	Stage 2	$64 \times 8 \times 25$	$[\ 1 \times 1$ conv, 3×3 conv $]\ \times 1$, s 2×2 $\begin{bmatrix} 3 \times 3 \text{ TA}, \\ 8 \times 3 \text{ conv}, \quad BiLSTM \end{bmatrix} \times 3$
	Stage 3	$128 \times 4 \times 25$	$[\ 1 \times 1$ conv, 3×3 conv $]\ \times 1$, s 2×1 $\begin{bmatrix} 3 \times 3 \text{ TA}, \\ 4 \times 3 \text{ conv}, \quad BiLSTM \end{bmatrix} \times 5$
	Stage 4	$256 \times 2 \times 25$	$[\ 1 \times 1$ conv, 3×3 conv $]\ \times 1$, s 2×1 $\begin{bmatrix} 3 \times 3 \text{ TA}, \\ 2 \times 3 \text{ conv}, \quad BiLSTM \end{bmatrix} \times 5$
	Stage 5	$512 \times 1 \times 25$	$[\ 1 \times 1$ conv, 3×3 conv $]\ \times 1$, s 2×1 $\begin{bmatrix} 3 \times 3 \text{ TA}, \\ 1 \times 3 \text{ conv}, \quad BiLSTM \end{bmatrix} \times 2$
Decoder	Att. LSTM	*	256 attention units 256 hidden units

TABLE I: Text recognition network configurations of DRNet (TA+NC version). "s" stands for stride of the first convolution layer in each stage. "Out Size" is feature map size output from each stage (channel $\times$ height $\times$ width). "*" means dynamic output length. "Att. LSTM" stands for attentional LSTM decoder [11].

$$e_{t,i} = W_e^T \tanh(W_f f_i + W_h h_{t-1} + b), \quad (4)$$

$$\alpha_{t,i} = \text{Softmax}(e_{t,i}), \quad (5)$$

$$g_t = \sum_i \alpha_{t,i} f_i, \quad (6)$$

$$y_t = \text{Softmax}(\text{RNN}((g_t, h_{t-1}, y_{t-1})), \quad (7)$$

where f represents the input feature of decoder, W_f, W_h, W_e are trainable weights, t is the time step of predicting one specific character, α is the attention weight, g is the context vector, h is the hidden state of RNN and y is the final prediction. Our DRNet can be end-to-end optimized according to

$$L = -\frac{1}{m} \sum_{t=1}^{m} \log\ p(y_t|I), \quad (8)$$

where m and I represents the number of characters and input image, respectively.

IV. Experiment

A. Datasets

Our proposed model is optimizd on synthetic datasets, Synth90K [57] and SynthText [58], without any finetuning on other datasets. To demonstrate the effectiveness of our DRNet, six scene text recognition datasets are employed to evaluate our method and we give a brief introduction about them in the following.

Synth90K(Sy90) [57] is a synthetic dataset containing 9 million word box images by rendering the commomly used words to background with some noise. It is widely used for training recognition models.

SynthText(ST) [58] is a synthetic text image dataset which originally created for scene text detection task and is extended to text recognition by cropping the text instance according to the given bounding boxes. The character-level and text-level bounding boxes in this dataset makes it possible for the supervised training process of our rectification module.

IIIT5K-Words (IIIT) [59] includes 2000 training images and 3000 testing images which are collected from Google image searches.

Street View Text (SVT) [60] contains 647 outdoor street images collected by Google Street View. Many images in this dataset are seriously corrupted by blur and low resolution.

ICDAR 2013 (IC13) [61] contains 1015 scene text images which are mostly inherited from IC03.

ICDAR 2015 (IC15) [62] contains 2077 texts images captured incidentally by Google Glasses.

SVT-Perspective (SVTP) [63] contains 645 testing images cropped from SVT dataset [60]. To evaluation recognition performance for perspective text instances, the heavily distorted word images are carefully selected from SVT dataset.

CUTE80 [64] contains 288 word images which focus on curved texts. The word images are cropped from 80 high resolution natural images in CUTE dataset.

B. Implement Details

For the fairly comparison, our DRNet keeps the same configurations as ASTER [11]. Specially, training samples will be resized to 64×256 as the input of TPS module [56] and then transformed to 32×100 before the feature extractor stage. The number of control points for TPS is 20, same as ASTER. Attentional LSTM decoder is used in prediction stage as the RNN-based one-dimensional decoder, where the number of attention units and hidden units are both 256. 94 character classes will be recognized in our model, including digits, 32 ASCII punctuation, upper and lower case letters.

We implement ADADELTA [65] with the hyper-parameters $\text{eps} = 10^{-6}$ and $\text{rho} = 0.9$ as our optimizer and the initial learning rate is set to $\text{lr}_{\text{initial}} = 1$ which will be divided by 10 at 3 and 5 epochs. We train the model with batch size 512 for 6 epochs. At test time, beam search is used which keeps the top-k candidates with the highest accumulative scores, and k is set to 5 in our experiments.

Our proposed model is optimized on synthetic datasets, Synth90K and SynthText, without any finetuning on other datasets. All extensive experiments are implemented on Pytorch platform with NVIDIA GeForce RTX 2080Ti graphic cards.

C. Ablation Study for Overall Pipeline

In this section, we conduct several ablation experiments to verify our assumption that extracting local visual and long-range contextual information simultaneously is more effective than sequentially. To begin with, we firstly explore the usage of contextual information with different configurations in STR task from the naive CNN model to our DRNet step by step. Then, we investigate the effect of the long-range contextual information at a low level, which verifies our assumption in previous section.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3171108, IEEE Transactions on Multimedia

methods	IIIT5k	SVT	IC13	IC15	SVTP	CUTE80
Vis	93.1	86.2	93.1	77.3	79.2	78.5
Vis+Cont (ASTER*)	93.2	89.2	91.0	78.0	81.2	81.9
Vis+Cont (DRNet)	**93.6**	**89.6**	**94.6**	**81.6**	**83.4**	**82.3**
Vis+Cont (cascad-DR)	93.2	88.9	94.0	81.2	82.0	80.2
Vis+Cont (stage-DR)	93.3	89.2	94.0	76.1	80.9	79.5

TABLE II: Ablation Study for Long-range Contextual Branch. Comparison of recognition accuracy. "Vis" and "Cont" represent extracting local visual and contextual feature respectively. ASTER* represents the re-implementation performance using officially released pytorch version code.

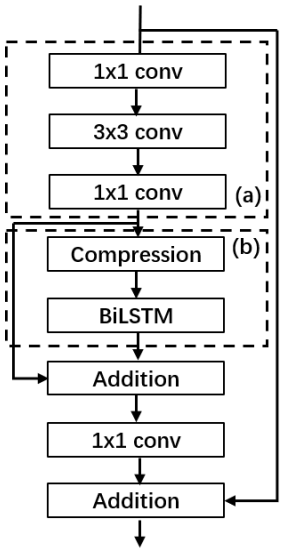


Fig. 7: Structure of cascade-DR-block. (a) represents the local visual branch and (b) represents the long-range contextual branch. Compression in this experiment are normal convolution (NC) version. Addition represents element wise addition. 3×3 convolutions are used for local visual branch in these experiments instead of TA operation.

1) : The Different Configurations of the Long-Range Contextual Branch

Five different models are developed for comparison in this subsection as shown in TABLE II. For fair comparison, we only use normal 3×3 convolution to extract local visual feature rather than TA operation in all of the following experiments. The first model, termed as "Vis", only employs local visual information to recognize scene text by removing sequential layer in ASTER. ASTER is the second model, termed as "Vis+Cont(ASTER*)", which extract local and contextual relations in a sequential manner. We then add our long-range contextual branch (NC version) to form the CONV + NC version of DR-block and extract block-wise contextual relation simultaneously (shown in Fig.8(a)), noted as "Vis+Cont (DRNet)". To perform a better comparison between extracting two relation simultaneously and sequentially, we design an cascaded-DR-block (shown in Fig .7) which exactly has the same FLOPS and parameters as proposed DR-block (CONC + NC). Since they both extract long-range contextual relation started from low-level features, the only difference between DR-block and cascaded-DR-block is the data flow in each block and we term this model as "Vis+Cont (cascad-DR)".

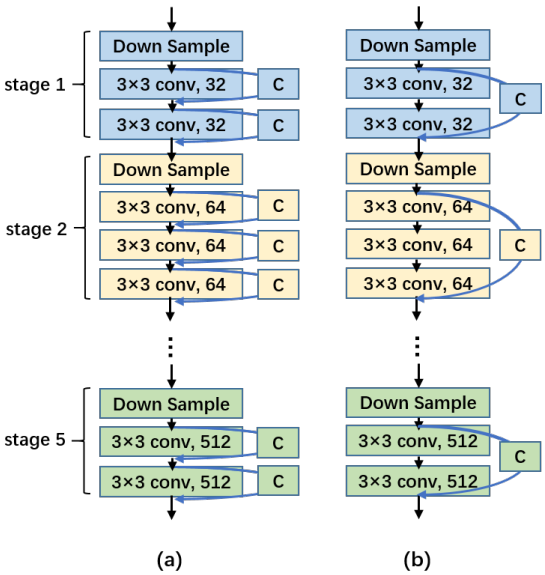


Fig. 8: Structure of "Vis+Cont (DRNet)" (a) and "Vis+Cont (stage-DR)" (b). Down Sample represents the first block in every stage which include an 1×1 convolution and a 3×3 convolution (details can be seen in TABLE I). C represents contextual branch (NC version) **instead of** residual link.

Moreover, to explore the linking relation of contextual branch, we change the block-wise contextual branch to stage-wise noted as "Vis+Cont (stage-DR)". In this model, the contextual branch calculates the relation at the start of each stage and adds to the end of entire stage, which is visualized in Fig. 8(b).

From the experimental results in TABLE II, we can draw the following conclusions: Firstly, the contextual information (in the second row) can efficiently model character relations and improve recognition performance, which has been reported in several studies [22]. Secondly, compared with "Vis+Cont (ASTER*)", "Vis+Cont (cascad-DR)" outperforms classic text recognition pipeline in 4 standard benchmarks, which proves that low-level global-contextual relation has positive contribution to text recognition. Thirdly, comparing "Vis+Cont (cascad-DR)" and "Vis+Cont (DRNet)", in which "Vis+Cont (DRNet)" has better performance in all 6 benchmarks, our assumption can be verified that processing local visual and long-range contextual relation simultaneously is better than sequentially. Finally, the drop of performance in "Vis+Cont (stage-DR)" indicates that block-wise contextual branch is better than stage-wise, since the features change severely after a whole stage.

2) : The Effect of Contextual Information at Low-level

To further verify the effectiveness of contextual information at a low level, extensive experiments for truncated model are further designed. Upon the trained model, all the trainable weights of encoder are frozen, and the output features of local visual branch or long-range contextual branch from each stage are passed to the non-frozen trainable decoder directly to perform the recognition as shown in Fig. 9. Then we train this truncated model and compare their average recognition

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3171108, IEEE Transactions on Multimedia

Variants	Stage 1	Stage 2	Stage 3	Stage 4
DR-local	76.6	81.8	85.8	87.6
DR-contextual	80.1	83.7	84.5	87.0

TABLE III: Comparison of average recognition performance using local or contextual information by truncated model of different stage. "DR-local" represents model using local information and "DR-contextual" represents model using contextual information.

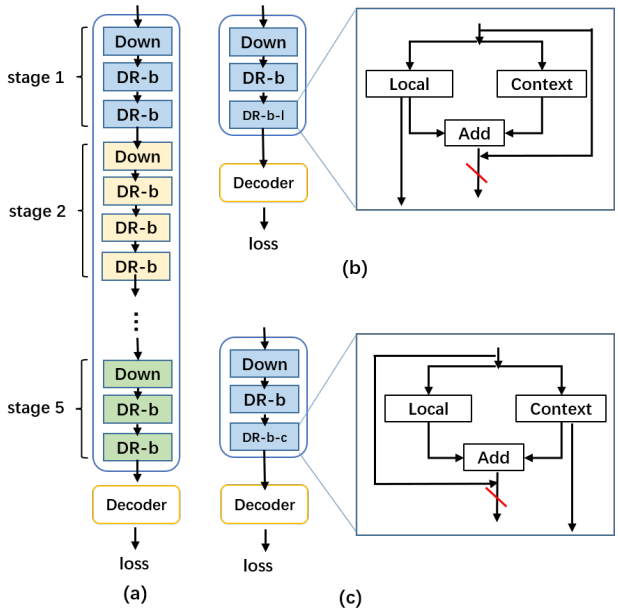


Fig. 9: Visualization of truncated model. (a) is the complete model pipeline which is not truncated. (b) is the visualization of model truncated on stage 1 and using local visual branch to predict. (c) is the visualization of model truncated on stage 1 and using contextual branch to predict. "Down" represents the block with down sample. "DR-b" represents DR-block. "DR-b-l" and "DR-b-c" represent modified DR-blocks which only output feature form local branch or contextual branch respectively. The simplified version of "DR-b-l" and "DR-b-c" (without a convolution operation and addition operation for better presentation, for detailed structure, please refer to Fig. 6) are shown in the right rectangle.

performance on the benchmarks. The result is shown in TABLE III and there are two findings: Firstly, the performance of these two models increase as the preserved stages increase, which is reasonable since the performance will grow as the model become deeper. Secondly, the performance of "DR-contextual" is higher than "DR-local" in low-level stage while lower in high-level stage. In the low-level stage, it is hard for local-visual features to make the right prediction while the long-range contextual has a better performance, which proves the feasibility for long-range contextual relation to facilitate the feature extraction at a low level. Therefore, this experiment proves that long-range contextual information at a low level is actually important though being long ignored.

D. Ablation Study for Topological-aware Operator

In this section, we explore the effect of our TA operator in local-visual branch with various settings.

Variants	IIIT5k	SVT	IC13	IC15	SVTP	CUTE80
DRNet(3 $\times$3)	93.7	90.6	95.8	81.6	83.6	83.0
DRNet(5 $\times$5)	94.0	89.9	95.4	81.5	82.4	81.9
DRNet(7 $\times$7)	93.7	89.6	95.7	81.2	83.2	82.2

TABLE IV: Comparison of recognition accuracy using different kernel size of TA-operator in local-visual branch.

Variants	IIIT5k	SVT	IC13	IC15	SVTP	CUTE80
TA-operator	**93.7**	**90.6**	**95.8**	**81.6**	**83.6**	**83.0**
Self-attention [41]	93.4	89.7	93.8	80.5	81.7	81.5

TABLE V: Comparison of recognition accuracy with self attention model.

1) : Analysis on Kernel Size

The experiments that using 5×5 and 7×7 size TA-operator are conducted and the experimental results are shown in TABLE IV. The performance of using large kernels is slightly lower than small kernels and we think this phenomenon is reasonable. There are at least two reasons: Firstly, our TA operation does not have position embedding. Without position embedding, the large kernel will cause more pixel loss there position information, which weaken the ability to extract salient features. Though causing the low performance of large kernels, position embedding should be discard to keep this rotation-invariant attribute. Secondly, large kernels tend to make the feature plain, which is the same as the drawbacks of using large convolution kernels.

2) : Comparison on the self-attention model

The TA operation is similar with the self-attention module. The key differences between TA operator and self-attention module is our TA operator extract the visual information in a local region, while most existing self-attention modules extract visual information from the whole feature map. In such case, we think the self-attention module cannot effectively model the intra-character relations, since the inter-character relations are inevitably introduced via the attention mechanism. We utilize the widely used self-attention module, Non-local [41] to verify this assumption. Due to the extremely high capacity on graphic card memory, it is not realistic to replace every TA-operator with non-local layer. Thus, we replace every last DR-block in each stage with new designed non-local based DR-block, where non-local layer is used for the local visual branch. In this way, there are 5 non-local layers are used in total, which is actually a commonly-used configuration in original paper [41]. The recognition accuracy is shown in TABLE V. Our model outperform Non-local in all benchmarks, which demonstrate the effectiveness of our TA operator.

E. Ablation study for DRNet

In this section, we perform extensive experiments to verify the design of our DR-block. For the local visual branch, we compare the TA-operation with the normal 3×3 Convolution operation. For the long-range contextual branch, in this paper, we propose two different approaches, normal convolution based (NC-based) and group convolution based (GC-based), to adaptively compress vertical information. The NC-based version directly extract and compress vertical information via a normal $h \times 3$ convolutional operation. For the better computation and memory efficiency, the GC-based approach reshapes the feature map by combining the vertical information as a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3171108, IEEE Transactions on Multimedia

Variants	IIIT5k	SVT	IC13	IC15	SVTP	CUTE80	FLOPS(G)	Parameters (M)	Training time(ms)	Inference time(ms)
TA only	92.6	89.0	91.9	76.4	78.9	79.9	3.0	11.3	2.67	13.1
ASTER*	93.2	89.2	91.0	78.0	81.2	81.9	3.6	14.8	1.23	8.2
TA + NC	**93.7**	**90.6**	**95.8**	**81.6**	**83.6**	**83.0**	4.8	25.5	2.83	14.9
TA + GC	93.2	90.3	94.0	81.2	81.4	82.3	4.1	20.6	2.81	18.1
TA + MAXP	93.2	88.1	94.7	81.0	79.7	81.2	3.7	25.4	2.73	13.6
CONV + NC	93.6	89.6	94.6	**81.6**	83.4	82.3	6.0	30.3	1.56	8.5
CONV + GC	93.3	89.3	94.3	81.1	82.6	81.6	5.3	25.4	1.77	9.8
CONV + MAXP	93.4	88.1	93.3	78.6	80.9	79.2	4.1	30.4	1.42	8.1

TABLE VI: Comparison of recognition accuracy with different configurations. "TA" and "CONV" represent using 3×3 TA operation and 3×3 convolution in local visual branch respectively. "NC" and "GC" represent using normal convolution version and group convolution version in long-range contextual branch respectively. Comparison of FLOPS, number of parameters, training time and inference time. ASTER* represents the re-implementation performance using officially released pytorch version code.

group, and then utilizes the group convolution to compress vertical information. We validate the efficient of different feature compression approaches by comparing the NC-based and GC-based approaches with common-used Max Pooling operation.

Experimental results are summarized in TABLE VI. The recognition accuracy of only-TA model is slightly lower than ASTER*, since no BiLSTM is utilized to extract long-range contextual information on top of the model in the only-TA model, which in term verify the importance of contextual information. With the help of long-range contextual branch, "TA+NC" significantly improves the recognition performance on every benchmark, and outperform the ASTER* in all benchmarks with a large margin. For local visual branch in DRNet, our TA operation is better than normal convolution ("CONV") due to the topological awareness discussed in early section. Especially for the NC version, TA has the best performance in all benchmarks and increases the accuracy of CUTE, SVT and IC13 by 0.7%, 1.0% and 1.4%. For long-range contextual branch in DRNet, max pooling strategy ("MAXP") has a lowest performance since it almost discards all vertical information while convolution based strategies ("NC" and "GC") adaptively retain the vertical relations. Moreover, the normal convolution ("NC") is better than group convolution ("GC") and achieves best performance, because the group convolution only models intra-group relations while normal convolution models both intra- and inter-group relations along channel dimension. Therefore, "TA + NC" has the best performance and we employ this configuration as our final DRNet model.

Moreover, we study the FLOPS, number of parameters, training and inference time for our proposed method. We test our model and baseline on one single RTX-2080Ti GPU and Intel(R) Xeon(R) CPU E5-2640 v3 with 2.60GHz. As shown in TABLE VI, although our model increase FLOPS and Parameters due to dual branch structure, our proposed method keeps real-time inference. Moreover, compared with "Conv + NC", we think the increased inference time mainly comes from that our TA operation is not well-optimized as the state-of-the-art convolutional operation. Specifically, we implement TA operation by using PyTorch's unfold function. Compared with the PyTorch official implementation of convolution and BiLSTM operations, the unfold implementation of TA operation is quite slow, which becomes the bottleneck of our DRNet. We further examined the processing time of a single layer for 3×3 convolution, 3×3 TA, and BiLSTM operation, respectively. The input size is $32 \times 4 \times 25$ for convolution and TA operation, while $32 \times 1 \times 25$ for BiLSTM, since vertical information is compressed before BiLSTM. The experiment shows our TA operation is $26\times$ slower than the normal convolution and $2\times$ slower then the BiLSTM. Moreover, the low-efficient of unfold implementation is also observed by the paper [55], and they provide two advanced implementations, the customized CUDA kernel approach and the sliding chunk approach namely, to solve this issue.

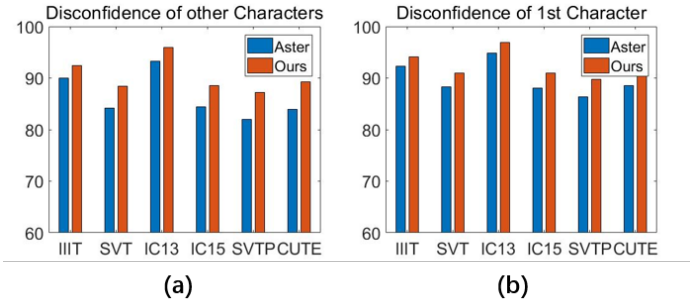


Fig. 10: Comparison of confidence between ASTER and our DRNet. (a) shows average confidence for the first characters and (b) shows average confidence for the latter characters.

Variants	IIIT5k	SVT	IC13	IC15	SVTP	CUTE80
SAR	91.5	84.5	91.0	69.2	76.4	**83.3**
DRNet-SAR	**94.4**	**88.6**	**93.8**	**76.6**	**77.4**	81.9

TABLE VII: Comparison of recognition accuracy with 2D attention model. Our DRNet-SAR outperforms original SAR in five of the six datasets which verifies the effectiveness in 2D recognition models.

F. Comparison with baseline

In this section, we compare our DRNet to the baseline model in different ways to prove the superiority of our model.

1) : Similarity

To illustrate the topological awareness of our TA operation, we compare the similarity of output features with different rotation angles.

We select all single-character images in CUTE80 dataset [64] and resize them to the shape of 64×64. Then we rotate these images with three angles $90°, 180°, 270°$, and pad them into 64×256 pixels, which is the input size of our DRNet. The rectification stage is removed and all images are sent to

TA / Conv	Origin	Rotate 90°	Rotate 180°	Rotate 270°
Origin	-	**68.3** / 59.6	**72.7** / 61.5	**66.9** / 59.1
Rotate 90	**68.3** / 59.6	-	**64.7** / 61.0	**72.9** / 66.7
Rotate 180	**72.7** / 61.5	**64.7** / 61.0	-	**69.6** / 62.8
Rotate 270	**66.9** / 59.1	**72.9** / 66.7	**69.6** / 62.8	-

TABLE VIII: Average cosine similarities of features by TA operation and normal convolution. "TA" and "Conv" represent TA operation and normal convolution, respectively. Our TA operation is more robust when extracting rotated images.

the trained feature extractor directly. Since three fourths of the input image are padded with zero, we only calculate the first one fourth of the output feature with cosine similarity. Experimental results are shown in TABLE VIII. Comparing with 3×3 normal convolution, our TA operation keeps a greater similarity in all rotating angles, which means our TA operation is more robust than normal convolution and can extract similar visual features for the same character.

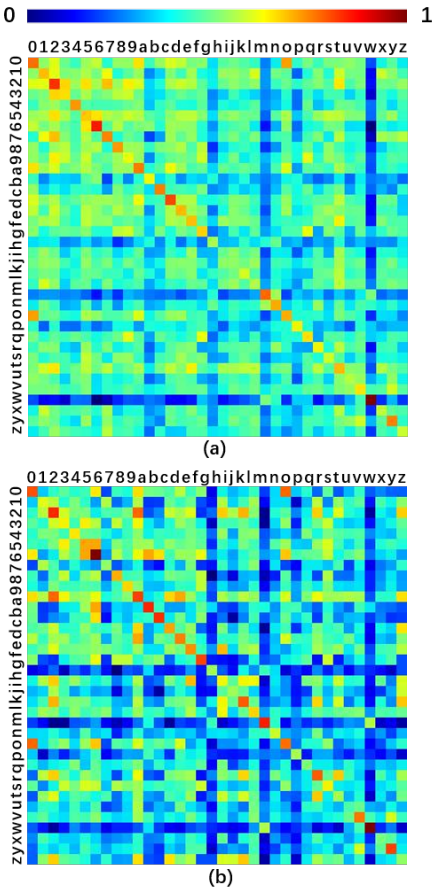


Fig. 11: The visualization of cosine similarity between characters of same or different categories. (a) is the similarity matrix of ASTER and (b) is of DRNet.

Moreover, we compare the cosine similarity of extracted feature by DRNet and ASTER in the same or different categories. Specifically, we randomly cut 100 character images for each category from Syntext [58] to form our image pool and make a 36×36 matrix to show the feature similarities between each class. In order to fill each element in the matrix, we further randomly selected 100 pairs of character images for each comparison to calculate their average cosine similarities. After getting the matrix, we normalize the matrix into the region $[0, 1]$, and then visualize these two 36×36 matrices using heatmap for better representation. The experimental results are shown in Fig. 11, where the characters with similar topological structures have higher similarities while characters with different topological structures have lower similarities. For example, the similarity between the character "o" and the number "0" is really high for both models. So we can conclude that our DRNet extracts more distinguishable features: higher score for the same class and lower for different classes.

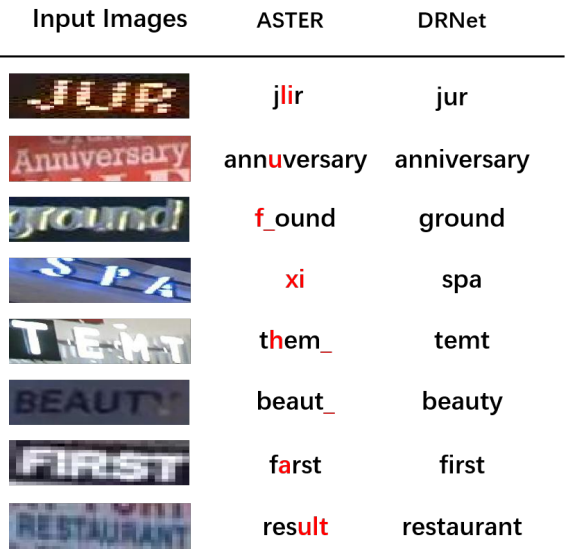


Fig. 12: Comparison of recognition results between ASTER and our DRNet. The first column shows the input images. Second column shows the predictions from ASTER where the wrong predictions are denoted in red color and the underlined position means the missing of characters. The third column shows the predictions from our DRNet.

2) : Confidence

To illustrate the effect of DRNet in reserving local visual and contextual information, we calculate the average confidence when text instances are correctly predicted, shown in Fig. 10. Since RNN-based decoder is mono-directional, the prediction of a latter character is based on the prediction of former ones. As a consequence, when predicting the first character, the dominant factor is local visual information while the later characters are influenced by both local and contextual factors. In Fig. 10(a), our DRNet (in orange) increases the confidence when predicting the first characters, indicating DRNet reserves more local visual information. In Fig. 10(b), our DRNet (in orange) increases the confidence of other characters in a larger extent compared with confidence of first characters, thus verifies our hypothesis that our DRNet extracts contextual information more effectively.

3) : Prediction

Fig. 12 visualizes the input images and predictions of baseline and our DRNet. As the introduction of long-range contextual relation in low-level feature, recognizer gains an

Methods	IIIT5k	SVT	IC13	IC15	SVTP	CUTE80
Mishra *et al.* [66]	64.1	73.2	-	-	-	-
Almazan *et al.* [67]	91.2	89.2	-	-	-	-
Yao *et al.* [68]	80.2	75.9	-	-	-	-
Gordo [69]	93.3	91.8	-	-	-	-
Jaderberg *et al.* [70]	-	80.7	90.8	-	-	-
Shi *et al.* [5]	81.9	81.9	88.6	-	71.8	59.2
Lee *et al.* [71]	78.4	80.7	90.0	-	-	-
Shi *et al.* [8]	81.2	82.7	89.6	-	-	-
Yang *et al.* [72]	-	-	-	-	75.8	69.3
Cheng *et al.* [29]	87.0	82.8	-	68.2	73.0	76.8
Liu *et al.* [73]	92.0	85.5	91.1	74.2	78.9	-
Bai *et al.* [74]	88.3	87.5	94.4	73.9	-	-
Liu *et al.* [75]	87.0	-	92.9	-	-	-
Liu *et al.* [76]	89.4	87.1	94.0	-	73.9	62.5
Liao *et al.* [30]	91.9	86.4	91.5	-	-	79.9
Shi *et al.* [11]	93.4	89.5	91.8	76.1	78.5	79.5
Luo *et al.* [77]	91.2	88.3	92.4	76.1	77.4	68.8
Zhan *et al.* [24]	93.3	90.2	91.3	76.9	79.6	83.3
Yang *et al.* [25]	94.4	88.9	93.9	78.7	80.8	87.5
Cheng *et al.* [78]	87.4	85.9	93.3	70.6	-	-
Zhang *et al.* [79]	88.3	88.6	93.7	78.7	-	76.3
Wang *et al.* [6]	91.5	84.5	91.0	69.2	76.4	83.3
Wang *et al.* [31]	94.3	89.2	94.2	74.5	80.0	84.4
Yue *et al.* [80]	**95.3**	88.1	-	77.1	79.5	**90.3**
Qiao *et al.* [81]	93.8	89.6	92.8	80.0	81.4	83.6
Gao *et al.* [82]	94.3	88.7	93.3	76.8	81.2	88.2
DRNet **(ours)**	93.7	**90.6**	**95.8**	**81.6**	**83.6**	83.0

TABLE IX: Comparison of recognition accuracy with State-of-the-art models.

overview of character distribution, thus misalignment problems are less likely to happen. For example, when predicting the first image "JUR" in Fig. 12, guided by the distribution information, the recognize will not regard the character "U" as combination of "L" and "I" but an individual character. Similar situation happens when predicting the image "BEAUTY" in Fig. 12, after knowing the character distribution, the recognizer is less likely to miss characters.

4) : 2D Attention Model

Above experiments show the effectiveness of DR-block in classic rectification based text recognition model. In this section we will demonstrate the generality of our model in 2D attentional model. SAR [6] is a text recognition model based on two-dimensional attention decoder and we replace the residual blocks in the model with our DR-block to form DRNet-SAR.

As shown in the TABLE VII, DRNet-SAR outperforms original SAR on 5 benchmarks with a large margin. The significant improvement over original SAR verifies the generality of our DRNet, which can be served as a plug-and-play module for any STR models.

G. Comparison with State-of-the-art

To demonstrate the effectiveness of our DRNet, we compare our model with state-of-the-art (SOTA) models on 6 public benchmarks. As shown in the TABLE IX, our DRNet achieves promising performance in 4 datasets of the six including SVT, IC13, IC15 and SVTP. Specifically, DRNet outperforms our baseline (ASTER) on all benchmarks and increases the recognition performance by 4.0%, 5.5%, 5.1% and 3.5% in IC13, IC15, SVTP and CUTE80 respectively. Furthermore, our model improves the recognition accuracy compared with SOTA by 1.6% and 2.2% in IC15 and SVTP datasets, which is not negligible and verifies our assumption that extracting local visual relation and long-range contextual relation simultaneously is better than sequentially. Moreover, without designing complex rectification or decoding module, we prove that the complete and effective representation for complementary relations can significant improve recognition performance of existing STR methods, which is ignored in recent works. Our DR-block is a effective plug-and-play feature extraction module and can be used as a direct replacement of convolutional block for any STR platforms to further boost recognition performance.

H. Limitations

We visualize the typical failure cases as shown in Fig. 13, which can be roughly divided into three categories. Firstly, characters in the image might be too similar to other characters and thus lead to the false prediction. For example, in Fig.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3171108, IEEE Transactions on Multimedia

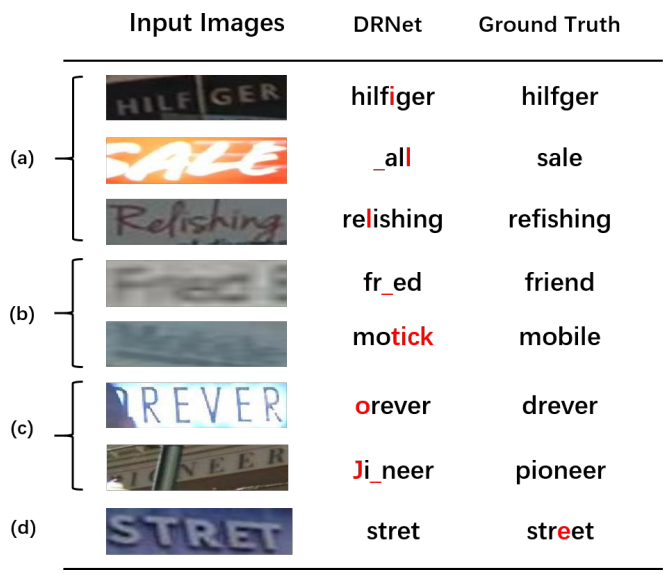


Fig. 13: Failure case. The first column shows the input images. Second column shows the predictions from our DRNet and the wrong predictions are denoted in red. Underlined position means the missing of characters. And the third column shows the ground truth. (a)(b)(c)(d) represents different kinds of errors.

13 (a) the "hilfger" image, the background that intruded into character region is too similar to the character "i" confuses the recognizer. Secondly, images might be too blury to be recognized as shown in Fig. 13 (b). Thirdly, characters might be sheltered by background and the remaining parts are difficult to recognize. As shown in Fig. 13 (c), the first character "d" is sheltered and impossible to be recognized. Finally, as shown in Fig. 13 (d), some ground truths in benchmarks are labeled falsely and thus cause the drop of accuracy.

Moreover, our model is not so effective on curved texts and long texts. Though the $h \times 3$ size convolution is used, the compressed feature with height 1 can not keep all the vertical information, which causes the relatively low performance on curved dataset. As for the long texts, our model is not difficult to predict for each single character. However, the width of our feature and the length of our contextual branch is only 25 pixel, which is not enough to model these texts exceeding 25 characters.

V. Conclusion

In this paper, we propose the dual relation module, which is a basic feature extractor to model visual local and long-range contextual relations in a paralleled two-branch structure. The local visual branch is designed for modelling intra-character relations and extracting discriminative features for different characters. Meanwhile, the long-range contextual branch is developed to compress vertical information and collect inter-character correlations. Moreover, we further formulate dual relation module into the dual relation block (DR-block), which can be served as a direct replacement of convolutional blocks in deep networks. To verify the effectiveness of our method, we develop our dual relation network (DRNet) by implementing our DR-block on classic scene text recognition platform. Experimental results demonstrate that our DR-block brings significant improvement on baseline model and achieves promising performance on a number of standard benchmarks.

References

[1] W. Wu, X. Chen, and J. Yang, "Incremental detection of text on road signs from video with application to a driving assistant system," in *ACM MM*, 2004.

[2] X. Chen, J. Yang, J. Zhang, and A. Waibel, "Automatic detection and recognition of signs from natural scenes," *IEEE Transactions on Image Processing*, vol. 13, no. 1, pp. 87–99, 2004.

[3] I. Kavati, G. K. Kumar, S. Kesagani, and K. S. Rao, "Signboard text translator: A guide to tourist," *Proceedings of International Journal of Electrical and Computer Engineering*, 2017.

[4] S. Karaoglu, R. Tao, J. C. van Gemert, and T. Gevers, "Con-text: Text detection for fine-grained object classification," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3965–3980, 2017.

[5] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4168–4176.

[6] H. Li, P. Wang, C. Shen, and G. Zhang, "Show, attend and read: A simple and strong baseline for irregular text recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8610–8617.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, 1997.

[8] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, 2016.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, 2014.

[11] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "Aster: An attentional scene text recognizer with flexible rectification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2035–2048, 2019.

[12] S. Zhang, Y. Liu, L. Jin, Z. Wei, and C. Shen, "Opmp: An omnidirectional pyramid mask proposal network for arbitrary-shape scene text detection," *IEEE Transactions on Multimedia*, vol. 23, pp. 454–467, 2021.

[13] M. Xue, P. Shivakumara, C. Zhang, Y. Xiao, T. Lu, U. Pal, D. Lopresti, and Z. Yang, "Arbitrarily-oriented text detection in low light natural scene images," *IEEE Transactions on Multimedia*, pp. 1–1, 2020.

[14] Y. Tang and X. Wu, "Scene text detection using superpixel-based stroke feature transform and deep learning based region classification," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2276–2288, 2018.

[15] X. Ren, Y. Zhou, J. He, K. Chen, X. Yang, and J. Sun, "A convolutional neural network-based chinese text detection algorithm via text structure modeling," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 506–518, 2017.

[16] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3111–3122, 2018.

[17] C. Yao, X. Bai, and W. Liu, "A unified framework for multioriented text detection and recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4737–4749, 2014.

[18] P. Dai, Y. Li, H. Zhang, J. Li, and X. Cao, "Accurate scene text detection via scale-aware data augmentation and shape similarity constraint," *IEEE Transactions on Multimedia*, 2021.

[19] Y. Liang and X. Li, "Reassembling shredded document stripes using word-path metric and greedy composition optimal matching solver," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1168–1181, 2020.

[20] J. Zhang, J. Sang, K. Xu, S. Wu, X. Zhao, Y. Sun, Y. Hu, and J. Yu, "Robust captchas towards malicious ocr," *IEEE Transactions on Multimedia*, vol. 23, pp. 2575–2587, 2021.

[21] 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.

[22] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, and H. Lee, "What is wrong with scene text recognition model comparisons? dataset and model analysis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[23] C. Bartz, H. Yang, and C. Meinel, "See: towards semi-supervised end-to-end scene text recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[24] F. Zhan and S. Lu, "Esir: End-to-end scene text recognition via iterative image rectification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2059–2068.

[25] M. Yang, Y. Guan, M. Liao, X. He, K. Bian, S. Bai, C. Yao, and X. Bai, "Symmetry-constrained rectification network for scene text recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9147–9156.

[26] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, vol. 28. Curran Associates, Inc., 2015.

[27] P. Dai, H. Zhang, and X. Cao, "Deep multi-scale context aware feature aggregation for curved scene text detection," *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 1969–1984, 2020.

[28] X. Wu, Q. Chen, Y. Xiao, W. Li, X. Liu, and B. Hu, "Lcsegnet: An efficient semantic segmentation network for large-scale complex chinese character recognition," *IEEE Transactions on Multimedia*, pp. 1–1, 2020.

[29] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, and S. Zhou, "Aon: Towards arbitrarily-oriented text recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5571–5579.

[30] M. Liao, J. Zhang, Z. Wan, F. Xie, J. Liang, P. Lyu, C. Yao, and X. Bai, "Scene text recognition from two-dimensional perspective," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8714–8721.

[31] T. Wang, Y. Zhu, L. Jin, C. Luo, X. Chen, Y. Wu, Q. Wang, and M. Cai, "Decoupled attention network for text recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 216–12 224.

[32] S. Fang, H. Xie, Y. Wang, Z. Mao, and Y. Zhang, "Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7098–7107.

[33] N. Nguyen, T. Nguyen, V. Tran, M.-T. Tran, T. D. Ngo, T. H. Nguyen, and M. Hoai, "Dictionary-guided scene text recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7383–7392.

[34] R. Yan, L. Peng, S. Xiao, and G. Yao, "Primitive representation learning for scene text recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 284–293.

[35] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, 1998.

[36] X. Saining, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017.

[37] C. François, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017.

[38] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017.

[39] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on Pattern Analysis and Machine Intelligence*, 2017.

[40] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[41] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018.

[42] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[43] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[44] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2020.

[45] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[46] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[47] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *arXiv preprint arXiv:1802.02611*, 2018.

[48] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[49] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 603–612.

[50] H. Ding, X. Jiang, B. Shuai, A. Qun Liu, and G. Wang, "Context contrasted feature and gated multi-scale aggregation for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2393–2402.

[51] J. Fu, J. Liu, Y. Wang, Y. Li, Y. Bao, J. Tang, and H. Lu, "Adaptive context network for scene parsing," in *Proceedings of the IEEE international conference on computer vision*, 2019.

[52] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157 vol.2.

[53] P. Dai, H. Zhang, and X. Cao, "Sloan: Scale-adaptive orientation attention network for scene text recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 1687–1701, 2020.

[54] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[55] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and J. Gao, "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2998–3008.

[56] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 6, pp. 567–585, 1989.

[57] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," *arXiv preprint arXiv:1406.2227*, 2014.

[58] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 2315–2324.

[59] A. Mishra, K. Alahari, and C. V. Jawahar, "Scene text recognition using higher order language priors," in *BMVC*, 2012.

[60] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2011, pp. 1457–1464.

[61] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. De Las Heras, "Icdar 2013 robust reading competition," in *International Conference on Document Analysis and Recognition*. IEEE, 2013, pp. 1484–1493.

[62] 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 *International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2015, pp. 1156–1160.

[63] T. Q. Phan, P. Shivakumara, S. Tian, and C. L. Tan, "Recognizing text with perspective distortion in natural scenes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 569–576.

[64] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, "A robust arbitrary text detection system for natural scene images," *Expert Systems with Applications*, vol. 41, no. 18, pp. 8027–8048, 2014.

[65] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," *CoRR*, 2012.

[66] A. Mishra, K. Alahari, and C. Jawahar, "Top-down and bottom-up cues for scene text recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2687–2694.

[67] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Word spotting and recognition with embedded attributes," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 12, pp. 2552–2566, 2014.

[68] C. Yao, X. Bai, B. Shi, and W. Liu, "Strokelets: A learned multi-scale representation for scene text recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4042–4049.
[69] A. Gordo, "Supervised mid-level features for word image representation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2015, pp. 2956–2964.
[70] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 1–20, 2016.
[71] C.-Y. Lee and S. Osindero, "Recursive recurrent nets with attention modeling for ocr in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2231–2239.
[72] X. Yang, D. He, Z. Zhou, D. Kifer, and C. L. Giles, "Learning to read irregular text with attention mechanisms," in *IJCAI*, 2017.
[73] W. Liu, C. Chen, and K.-Y. Wong, "Char-net: A character-aware neural network for distorted scene text recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
[74] F. Bai, Z. Cheng, Y. Niu, S. Pu, and S. Zhou, "Edit probability for scene text recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1508–1516.
[75] Z. Liu, Y. Li, F. Ren, W. L. Goh, and H. Yu, "Squeezedtext: A real-time scene text recognition by binary convolutional encoder-decoder network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
[76] Y. Liu, Z. Wang, H. Jin, and I. Wassell, "Synthetically supervised feature learning for scene text recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 435–451.
[77] C. Luo, L. Jin, and Z. Sun, "Moran: A multi-object rectified attention network for scene text recognition," *Pattern Recognition*, 2019.
[78] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, and S. Zhou, "Focusing attention: Towards accurate text recognition in natural images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5076–5084.
[79] Y. Zhang, S. Nie, S. Liang, and W. Liu, "Robust text image recognition via adversarial sequence-to-sequence domain adaptation," *IEEE Transactions on Image Processing*, vol. 30, pp. 3922–3933, 2021.
[80] X. Yue, Z. Kuang, C. Lin, H. Sun, and W. Zhang, "Robustscanner: Dynamically enhancing positional clues for robust text recognition," *European Conference on Computer Vision*, 2020.
[81] Z. Qiao, Y. Zhou, D. Yang, Y. Zhou, and W. Wang, "Seed: Semantics enhanced encoder-decoder framework for scene text recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
[82] Y. Gao, Y. Chen, J. Wang, and H. Lu, "Progressive rectification network for irregular text recognition," *Science China Information Sciences*, vol. 63, no. 2, pp. 1–14, 2020.

Ming Li received the B.S. degree from Xi'an Jiaotong University, China, in 2020 and served as a Research Assistant in ShenZhen Key Lab of Computer Vision and Pattern Recognition, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, from 2019 to 2021. He is currently a Master student at Texas A&M university, US. His research interests include scene text detection, scene text recognition and Natural Language Processing.

Bin Fu is currently an Assistant Research Fellow with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. He received the Ph.D. degree from the University of Hong Kong, in 2006 and B.E. degree from Lanzhou University, in 2014. His research interests include semantic segmentation and scene text recognition.

Han Chen received the B.S. degree from Huazhong University of Science and Technology, China, in 2017. She is currently a Master student at Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences.

Junjun He is currently a Research Assistant with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences and Shanghai AI Laboratory. His research interests include computer vision and medical image computing, especially on dense prediction, multi-view learning, multi-modal learning and efficient model design. He has published more than 10 papers in international journals and conferences, including T-PAMI, TMI, CVPR, ICCV, ECCV, MICCAI, ISBI etc.

Yu Qiao is a professor with the Shenzhen Institutes of Advanced Technology (SIAT), the Chinese Academy of Science and Shanghai AI Laboratory. His research interests include computer vision, deep learning, and bioinformation. He has published more than 240 papers in international journals and conferences, including T-PAMI, IJCV, T-IP, T-SP, CVPR, ICCV etc. His H-index is 67, with 28,000 citations in Google scholar. He is a recipient of the distinguished paper award in AAAI 2021. His group achieved the first runner-up at the ImageNet Large Scale Visual Recognition Challenge 2015 in scene recognition, and the winner at the ActivityNet Large Scale Activity Recognition Challenge 2016 in video classification. He served as the program chair of IEEE ICIST 2014.