Look Back Again: Dual Parallel Attention Network for Accurate and Robust Scene Text Recognition

Zilong Fu

University of Science and Technology of China
Hefei, China
JeromeF@mail.ustc.edu.cn

Guoqing Jin

State Key Laboratory of Communication Content
Cognition, People's Daily Online
Beijing, China
jinguoqing@people.cn

Hongtao Xie*

University of Science and Technology of China
Hefei, China
htxie@ustc.edu.cn

Junbo Guo

State Key Laboratory of Communication Content
Cognition, People's Daily Online
Beijing, China
guojunbo@people.cn

ABSTRACT

Nowadays, it is a trend that using a parallel-decoupled encoder-decoder (PDED) framework in scene text recognition for its flexibility and efficiency. However, due to the inconsistent information content between queries and keys in the parallel positional attention module (PPAM) used in this kind of framework (queries: position information, keys: context and position information), visual misalignment tends to appear when confronting hard samples (e.g., blurred texts, irregular texts, or low-quality images). To tackle this issue, in this paper, we propose a dual parallel attention network (DPAN), in which a newly designed parallel context attention module (PCAM) is cascaded with the original PPAM, using linguistic contextual information to compensate for the information inconsistency between queries and keys. Specifically, in PCAM, we take the visual features from PPAM as inputs and present a bidirectional language model to enhance them with linguistic contexts to produce queries. In this way, we make the information content of the queries and keys consistent in PCAM, which helps to generate more precise visual glimpses to improve the entire PDED framework's accuracy and robustness. Experimental results verify the effectiveness of the proposed PCAM, showing the necessity of keeping the information consistency between queries and keys in the attention mechanism. On six benchmarks, including regular text and irregular text, the performance of DPAN surpasses the existing leading methods by large margins, achieving new state-of-the-art performance. The code is available on <https://github.com/Jackandrome/DPAN>.

CCS CONCEPTS

• Computing methodologies → Computer vision tasks.

*Corresponding Author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMR '21, August 21–24, 2021, Taipei, Taiwan

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8463-6/21/08...\$15.00

<https://doi.org/10.1145/3460426.3463674>

KEYWORDS

attention mechanism; text recognition; language model

ACM Reference Format:

Zilong Fu, Hongtao Xie, Guoqing Jin, and Junbo Guo. 2021. Look Back Again: Dual Parallel Attention Network for Accurate and Robust Scene Text Recognition. In *Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR '21), August 21–24, 2021, Taipei, Taiwan*. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3460426.3463674>

1 INTRODUCTION

Reading text from images or videos has been a hot research topic in the computer vision community for many years because of its numerous applications [32–34, 39]. With the development of sequence recognition, many effective text recognition methods have been proposed, making significant progress in this field. Despite this, recognizing texts in the wild scene still remains challenging for the large variations of text font, color, layout, and image quality.

Most current methods with remarkable performance follow the attention-based encoder-decoder architecture, trying to align and recognize character regions adaptively through various attention mechanisms. The early works [3, 26] mainly employ a sequential-coupled encoder-decoder (SCED) framework. However, due to the inherent characteristics of SCED (sequential decoding and coupled attention calculation), the visual misalignment is inevitable [31]. Besides, SCED is hard to be paralleled in training and inference, which is time-consuming and inefficient in the real application. The latest methods [19, 38] utilize the parallel-decoupled encoder-decoder (PDED) framework to alleviate the above problems. Typically, within the encoder of PDED, a parallel position attention module (PPAM) is implemented, taking the manual-defined character index and convolutional features as queries and keys to directly generate N visual glimpses. Even though PPAM avoids the accumulation and propagation of errors due to the module coupling and sequential decoding, the visual misalignment still exists in recognizing some hard samples, which limits PDED's robustness. As shown in Fig. 1, the attention regions are diffused when recognizing blurry text, irregular text, and text with crowded characters.

We suppose that the reason for the insufficient ability of PPAM lies in the information inconsistency between queries and keys when conducting attention operation. As proved by method [40],

Figure 1: The examples of visual misalignment caused by PPAM. The original images and the visualization of attention maps for 4 decoding slots are illustrated.

the keys and queries have the same information content in the vanilla attention mechanism, and both contain position and context information. The position information provides character localization guidance for the attention mechanism, and the context information provides word-specific knowledge which is conducive to accurately focus on a certain character region in a word. In comparison, PPAM merely uses the position embeddings as queries. These position embeddings encode statistical information about character positions throughout the entire training set and remain fixed and word-agnostic during inference. When dealing with hard samples outside of the source domain distribution, PPAM can neither maintain sensitivity to the position of the character region nor focus on specific characters in specific words, which makes its adaptability to various scenes reduced in real applications.

In this paper, we try to alleviate the visual misalignment by using a cascading architecture of two attention modules. The first attention module applies positional queries to produce coarse visual glimpses and the second attention module exploits them to generate contextualized queries for the following attention score calculation to achieve better visual feature alignment, just like a person looking at the image twice, once to confirm the rough character positions, the second time to accurately focus on the character areas. To this end, we develop a dual parallel attention network (DPAN), in which a newly designed parallel contextual attention module (PCAM) is presented to generate precise visual glimpses. Specifically, PCAM consists of two parts, the contextualization part and the attention part. For the contextualization part, we propose a bidirectional language model (BLM) to capture the linguistic contexts from the coarse visual glimpses generated by PPAM and consider those as the contextualized queries for the following attention process. For the attention part, we take the contextualized queries and the convolutional features as the input and apply scaled dot-product attention to generate the accurate visual glimpses. By attentioning twice, most visual misalignments on hard samples can be eliminated, which greatly improves the entire PDED framework's accuracy and robustness. Extensive ablation studies confirmed our analysis as mentioned above and the effectiveness of the proposed PCAM. Furthermore, we evaluate DPAN on 6 benchmarks, and the superior performance of DPAN demonstrates its effectiveness and robustness.

The contributions of our work can be summarized as follows: (1) We analyze the shortcomings of the parallel positional attention module (PPAM) commonly used in the parallel-decoupled encoder-decoder (PDED) framework and propose an effective parallel contextual attention module (PCAM) to mitigate the visual misalignment caused by it; (2) In PCAM, we introduce a bidirectional

language model (BLM) to contextualize the queries, which provides a practical approach to enhance query vectors for attention-based methods; (3) The proposed dual parallel attention network (DPAN) achieves new state-of-the-art performance both on regular and irregular benchmarks and performs well on hard samples (85.5% on ICDAR2015, 89.0% on SVTP and 91.9% on CUTE80)

2 RELATED WORK

The attention mechanism in text recognition is used for adaptive visual feature alignment (also called soft alignment). According to whether the visual feature alignment is performed in the network, we roughly divide the previous methods into two categories: attention-free methods and attention-based methods.

2.1 Attention-free Methods

Attention-free methods do not need to pay special attention to the salient features of character regions and can usually achieve efficient parallel recognition, such as CTC-based methods and segmentation-based methods. Most CTC-based methods use CNN to transform an image as sliced visual features, then use RNN to do the sequential feature modeling, and finally employ a CTC[4] decoder for parallel character prediction[7, 25]. To cope with the images with irregular text layout, Start-Net [18] proposes to apply STN [11] to rectify the text at the beginning of the feature extraction. Wan *et al.*[29] present a 2D-CTC formulation which adds the height dimension for path searching. Despite the high efficiency in the training and inference for CTC-based methods, the lack of visual feature alignment mechanism limits their capacities. The segmentation-based methods aim to accurately segment each character in the image and then classify it, being not affected by the text layout[15, 28]. From another perspective, the segmentation-based methods are to hard align the visual features by the means of localization. Although these methods perform well in the recognition of regular and irregular texts, they need character-level annotations for training, which greatly limits their practicality.

2.2 Attention-based Methods

2.2.1 methods with SCED framework. The early attention-based text recognition methods are inspired by the sequential-coupled encoder-decoder(SCED) framework, performing well in recognizing regular text. To deal with the irregular texts, some methods [26, 36, 41] propose to modify the text layout through various rectification modules. And some methods [14, 37] propose to advance the attention mechanism to align visual features in 2D space. To improve the robustness and accuracy of attention-based methods with SCED framework, FAN [3] designs a subnetwork to mitigate the “attention drift” problem. RobustScanner[40] propose a parallel position enhancement branch to solve the visual misalignment led by the diminishing positional information in queries. Plugnet [35] develops a flexible super-resolution unit to cope with the low-quality text images. Because of the SCED framework's sequential decoding nature, it is not efficient when processing long text and causes visual misalignment problems due to the accumulation of errors.

2.2.2 methods with PDED framework. To achieve accurate and efficient text recognition, latest methods apply parallel-decoupled

Figure 2: The architecture of Dual Parallel Attention Network (DPAN)

encoder-decoder(PDED) framework. DAN [31] decouples the encoder-decoder and proposes a parallel attention module to solve the visual misalignment in the SCED framework. Iyu *et al.* [19] propose the PPAM which uses randomly initialized vectors indicating positions as queries to generate aligned visual features in parallel. Based on [19], SRN [38] introduces stacked transformer encoder units to model the global linguistic contexts from visual features and fuses the linguistic information and visual information in decoding, achieving the new state-of-the-art performance. However, due to the lack of contextual information in queries in PPAM, visual misalignment still exists when recognizing hard samples. The proposed DPAN in this paper aims to solve this problem by introducing a new parallel attention module that uses linguistic semantics to generate contextualized query vectors for better visual feature alignment.

3 METHOD

3.1 Network Architecture

The pipeline of the proposed DPAN is illustrated in Fig.2, which consists of four components: visual feature extractor, parallel positional attention module (PPAM), parallel contextual attention module (PCAM), and parallel decoder. Given a text image x , the visual feature extractor firstly transforms x to the 2D features $V \subseteq \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times c}$, where H and W are the size of x and c is the feature dimension. Here, we follow other works [19, 38], employing Resnet[6] as the CNN backbone and the stacked transformer encoder units [27] to enhance the visual features. Then the PPAM is used to generate the coarse visual glimpses G_1 . After that, PCAM takes the projections of G_1 to produce contextualized queries Q_c to generate the precise visual glimpses G_2 . In the end, G_2 and Q_c are fed into the parallel decoder to predict characters. In practice, we implement the gated fusion unit in [38] to do the dynamic feature fusion of G_2 and Q_c .

3.2 Parallel Positional Attention Module

Parallel positional attention module (PPAM) is firstly used in [19] demonstrating its ability to simultaneously focus on multiple specific areas in a two-dimensional space. In DPAN, we apply the initial version of PPAM to generate coarse visual glimpses. The process can be formulated as:

$$G_1 = \text{softmax}\left(\frac{Q_p K^T}{\sqrt{c}}\right)V \quad (1)$$

, where $K = V \subseteq \mathbb{R}^{l \times c}$ ($l = \frac{H}{4} \times \frac{W}{4}$) are flattened features from visual feature extractor, $Q_p \subseteq \mathbb{R}^{N \times c}$ indicates the embedding vectors of the manual-defined character index(maxlength is N), and $G_1 \subseteq \mathbb{R}^{N \times c}$ is the focused N character regions.

K and V have absolute position information and global context information as they are processed by the transformer units. What does Q_p encode during the training? For this problem, DETR [2] conducted experiments and found that each positional query shows strong sensitivity to objects' position rather than specific objects' semantic classes. And its sensitivity is highly related to the object distribution in the training set. Therefore, we analyze that in the training of PPAM, Q_p encodes more statistical information about the character position in the training set, which makes it more sensitive to character position rather than character category during inference. When encountering hard samples outside the training set distribution, the position-sensitivity of Q_p is greatly reduced, and visual misalignment tends to appear.

3.3 Parallel Contextual Attention Module

Figure 3: The architecture of Bidirectional Language Model

To alleviate the visual misalignment brought by the deficiencies of PPAM, we propose a new parallel contextual attention module (PCAM) in which we maintain the information content between queries and keys consistent. As shown in Fig.2, PCAM consists of two parts: a bidirectional language model used to produce contextualized queries Q_c and an attention operation used to generate new visual glimpses G_2 .

3.3.1 Bidirectional Language Model. We firstly review the vanilla attention mechanism exploited in the SCED framework to

figure out what kind of information is apt to contextualize query vectors. The formulation of vanilla attention is as follows:

$$\alpha_t = \text{softmax}(h_t^T V'') \quad (2)$$

$$h_t = F_{seq}(s_t, h_{t-1}), s_t = \begin{cases} F_{emb}(y_{t-1}), & t > 1 \\ \langle \text{START} \rangle, & t = 1 \end{cases} \quad (3)$$

, where α_t indicates the attention scores computed at t time step, h_t is the query and V'' is the visual features from the encoder. In Eq.2, y_{t-1} is the predicted character at the former time step, s_t represents its embedding vector, and F_{seq} is a RNN-like sequence modeling module, which encodes the relative position information and the context information for h_t . Here, we use F to denote the overall encoding process for all the past time steps, and the query can be reformulated as: $h_t = F(h_0, l(y_1, y_2, y_3, \dots, y_{t-1}))$, $t > 1$, where $l(\cdot)$ takes the previous predicted characters as input, modelling the unidirectional linguistic context for each query h_t ($t > 1$).

Figure 4: (a): lc_4 is the bidirectional linguistic contexts of the fourth visual feature in G_2 when the attention mask is applied. The squares with different colors represent different weights. (b): The workflow of BLM. q_4 is the contextualized query. An auxiliary loss is introduced to facilitate linguistic context modeling.

Hence, in PCAM, we also resort to linguistic contextual information to contextualize G_1 as queries. But unlike SCDE, PDED is not limited by sequential decoding and has the potential to utilize bidirectional linguistic context. To this end, we propose a bidirectional language model (BLM). The architecture of BLM is illustrated in Fig.3. Generally, BLM is the variant of stacked original transformer encoder units where the multi-head self-attention is replaced with multi-head cross-attention to prevent the information leakage across positions. Within the multi-head cross-attention layer, we take the linear projection of G_1 with positional encodings as keys and the feature indexes' position embeddings as queries. The reason why we only use positional queries here is that in the word-level linguistic context modeling, we concern more about the aggregation of features rather than their specific combinations. In addition, an attention mask is introduced in the cross-attention computing to avoid “seeing itself” for more effective context learning (shown in Fig.4(a)). Overall, the processing within a BLM layer can be formulated as:

$$K' = V' = \text{Pos}[\text{Pro}(G_1)] \quad (4)$$

$$S_l = F'(\text{softmax}(\frac{Q_l K'^T}{\sqrt{c}} + M)V') \quad (5)$$

, where S_l is the output of the l -th layer of BLM, the $\text{pro}(\cdot)$ is the linear projection layer that projects G_1 from visual space to linguistic space, $\text{Pos}[\cdot]$ denotes the position encoding in Transformer, and $M \subseteq \mathbb{R}^{N \times N}$ ($m_{ij} = 0$, if $i \neq j$; $m_{ij} = -\infty$, if $i = j$) is the attention mask matrix. By repeating the above process N times we got the contextualized queries Q_c . An example in Fig.4(b) qualitatively illustrates the workflow of BLM. It is worth noting that in DPAN, the gradient propagation between BLM and PPAM is stopped during the training phase. This is because we regard language modeling and visual feature extraction as two separate tasks which have different solution spaces and different convergence speeds. Therefore, we enforce PPAM and BLM to learn their specialized knowledge independently.

3.3.2 Attention Operation. In PCAM, we still adopt scaled dot-product attention. But different from PPAM, the queries and keys both contain position and context information, which ensures the quality of the G_2 . The process can be formulated as:

$$G_2 = \text{softmax}(\frac{\text{Pos}[Q_c]K^T}{\sqrt{c}}V) \quad (6)$$

, where $\text{Pos}[\cdot]$ indicates the position encoding function, and $K = V$ are the same as those in PPAM. $G_2 \subseteq \mathbb{R}^{N \times c}$ and $Q_c \subseteq \mathbb{R}^{N \times c}$ are fed to parallel decoder to form the fused features $F_c \subseteq \mathbb{R}^{N \times c}$ which are used to parallelly predict characters

3.4 Optimization

The proposed DPAN is an end-to-end trainable network with the training objectives as:

$$L = \lambda_1 l_{G_1} + \lambda_2 l_{Q_c} + \lambda_3 l_F \quad (7)$$

where l_{G_1} , l_{Q_c} , and l_F are cross-entropy loss for PPAM, PCAM and parallel decoder respectively. The parameters λ_1 , λ_2 , and λ_3 are balance factors, which are set to 1 during training.

4 EXPERIMENT

4.1 Dataset

For the fair comparison with other works, we follow their training settings that only use two synthetic text datasets MJSynth (MJ) [9, 10] and SynthText (ST) [5] to train DPAN. To evaluate the performance of the proposed model in various scenes, we choose 6 public benchmarks as the evaluation set, including 3 regular text datasets (ICDAR2013(IC13) [13], Street View Text(SVT) [30], IIIT5K-words(IIIT) [21]) and 3 irregular text datasets (ICDAR2015(IC15) [12], CUTE80 (CUTE) [24], Street View Text Perspective(SVTP) [22]). The detailed description of these datasets can refer to the previous work [38].

4.2 Implementation Details

All The feature dimensions c after the visual feature extractor are set to 512. The max length of character index and the decoding length N is set to 25. There are 8 attention heads contained in each transformer unit in visual feature extractor and BLM, respectively. But the exact stacked layers of transformer unit within these two

Table 1: Ablation study of PCAM. DF is short for decoded features. OA. is short for overall accuracy

Model	DF	IC13	IIIT	SVT	IC15	SVTP	CUTE	OA
Base Model	G_1	95.1	94.7	90.0	81.9	83.4	85.4	89.9
	F	95.9	95.3	92.6	83.3	85.4	86.9	91.0
DPAN	G_1	94.4	94.9	89.5	80.7	83.3	86.5	89.4
	G_2	96.4	95.4	92.7	83.0	87.8	87.8	91.2
	F	97.1	96.0	92.6	84.5	87.4	88.3	91.8

parts are super-parameter to be settled down in the following ablation studies. In the training stage, we adopt ADMD optimizer and set the initial learning rate as $1e^{-4}$, which is decayed to $1e^{-5}$ after 5 epochs. Moreover, some data augmentation operations like rotation, affine, perspective distortion, and color jitter, etc., are randomly added to the training images, and all the training images are resized to 128×32 . All the models are trained with the batch size of 384 on a workstation with 4 NVIDIA 2080Ti GPUs.

4.3 Ablation Study

4.3.1 Discussion on PCAM. To demonstrate the effectiveness of PCAM, we remove the attention operation within PCAM to form a baseline model for comparison. The number of layers of transformer unit in visual feature extractor and BLM are set to 3 and 4 for these two models. Moreover, in the decoding stage, we take the output of PPAM, PPCM and gated fusion unit respectively to do the evaluation. Note that in the base model, F is the fusion of G_1 and the query vector Q_c . The results are shown in Tab.1, where the overall accuracy refers to the model’s performance when the data in the 6 data sets are mixed for evaluation. From the results of the second and last rows in Tab.1, equipping the base model with PCAM can get accuracy promotion on each dataset, especially on irregular text (1.4% promotion for CUTE and 2.0% promotion for SVTP), which proves the effectiveness of introducing PCAM into the PDED framework and the necessity of maintaining the consistency of information content between queries and keys in the attention operation. Besides, Tab.1 also shows the applicability of different information in decoding. For example, in the base model and DPAN, the fusion of visual information and linguistic context can obtain more accurate character prediction, which is also consistent with the experimental conclusion of SRN[38]. The experimental results of Tab.1 also indicates that using G_2 for decoding is better than using G_1 , which further proves that PCAM can alleviate the visual misalignment caused by PPAM. The qualitative comparison of G_1 and G_2 is shown in Fig.5

4.3.2 Discussion on BLM. We conduct several sets of ablation experiment to discuss the configuration of BLM. The original BLM introduced in Sec.3.3.1 is considered as the baseline named Base-BLM. To prove the rationality of its design, we compare it with four variants: BLM with gradient flow to PPAM(BLM with GF.), BLM output without no position encoding(BLM w/o PosInfo.), BLM output with position embedding(BLM with PosEmb.), and BLM with multi-head self-attention(BLM with self-attn.). The transformer layers in visual feature extractor and BLM remain the same settings with Sec. 4.3.1.

Table 2: Ablation study on the configuration of BLM

Model Cfg.	IC13	IIIT	SVT	IC15	SVTP	CUTE	OA
Base-BLM	97.1	96.0	92.6	84.5	87.4	88.3	91.8
BLM with GF	95.6	94.8	90.6	83.5	86.2	86.8	90.4
BLM w/o PosInfo.	96.3	96.2	93.7	83.8	87.6	87.5	91.8
BLM with PosEmb.	96.5	95.5	93.5	84.0	87.0	89.2	91.6
BLM with self-attn.	97.4	94.6	90.7	83.7	86.4	85.5	90.8

Table 3: Ablation study on performance boost. Trm. is short for transformer units

PT	Trms.	IC13	IIIT	SVT	IC15	SVTP	CUTE	OA
PPAM	2	96.6	95.5	93.4	83.0	86.5	87.8	91.2
	3	97.4	95.5	93.7	85.3	89.0	89.9	92.2
	4	96.7	96.0	94.6	84.7	87.9	91.0	92.2
BLM	3	96.3	95.4	92.0	84.0	87.3	87.8	91.4
	4	96.7	95.5	92.3	83.6	87.4	89.6	91.5
	5	96.3	95.2	93.0	83.3	86.8	86.1	91.1
	6	96.4	95.7	93.4	84.3	88.1	87.2	91.5
PPAM BLM	3,4	97.3	95.9	93.5	84.9	88.2	91.2	92.3

The first set of experiments is to verify the effectiveness of the gradient stopping strategy. Seen from the results of in Tab.2, the performance of Base-BLM on each dataset outperforms that of BLM with GF., indicating that stopping the gradient flow between BLM and PPAM is indeed conducive to the learning of two different tasks, that is linguistic context modeling and visual feature extraction. The second set of experiments is to find the best way to add positional information to Q_c . The results show that using positional encoding is better than using position embedding, but the presence or absence of position encoding has no effect on the overall performance of the model. We analyze the reason is that Q_c has already obtained the positional information from the keys during the process of linguistic context modeling. The third set of experiments is to study whether the “information leakage” will impact on BLM when self-attention is used. The information leakage here refers to the fact that features in different locations indirectly see themselves from the Queries during attention, which may help the network to cheat in context modeling. The comparison of the experimental results confirms our hypothesis. The overall accuracy of BLM with self-attn. is 1% lower than that of Base-BLM.

4.3.3 Discussion on further performance boost. In this section, we pre-train PPAM and BLM respectively to see if it leads to performance improvements. Meanwhile. We also study the impact of the stacked transformer units with different layers on the model’s performance. Note that when pre-training PPAM, the number of transform units in BLM is 4, and when pre-training PCAM, the number of transform units in visual feature extractor is 3. By comparing the experimental results in Tab.3 with the base model in Tab.1, we find that: (1) pre-training both PPAM and BLM is more

Table 4: Accuracy comparisons. Anno. is short for annotation. SA indicates synthadd [29]. wiki indicates wiki-103

	Method	Training data	Anno.	Regular Text			Irregular Text		
				IC13	SVT	IIIT5K	IC15	SVTP	CUTE80
CTC-based	2019 Wan <i>et al.</i> [2D-CTC] [29]	ST	word	93.9	90.0	94.7	75.5	79.2	81.3
	2020 Hu <i>et al.</i> [GTC] [8]	MJ+ST+SA	word	94.4	92.9	95.8	80.4	85.5	92.0
Segmentation-Based	2020 Wan <i>et al.</i> [TextScanner] [28]	MJ+ST	word, char	92.9	90.1	93.9	79.4	84.3	83.3
	2019 Liao <i>et al.</i> [17]	MJ	word, char	91.5	86.4	91.9	-	-	79.9
SCED-based	2019 Liao <i>et al.</i> [SAM] [16]	MT+ST	word	95.3	90.6	93.9	77.3	82.2	87.8
	2020 Yue <i>et al.</i> [RobustScanner] [40]	MJ+ST	word	94.8	88.1	95.3	77.1	79.5	90.3
	2020 Qiao <i>et al.</i> [SEED] [23]	MJ+SJ	word	92.8	89.6	93.8	80.0	81.4	83.6
	2019 bartz <i>et al.</i> [1]	MJ+ST	word	93.1	89.2	94.6	74.2	83.1	89.6
	2020 Yan <i>et al.</i> [PlugNet] [35]	MJ+ST	word	95.0	92.3	94.4	82.2	84.3	85.0
PDED-based	2019 Lyu <i>et al.</i> [19]	MJ+SJ	word	92.7	90.1	94.0	76.3	82.3	86.8
	2020 Wang <i>et al.</i> [DAN] [31]	MJ+SJ	word	93.9	89.2	94.3	74.5	80.0	84.4
	2020 Yu <i>et al.</i> [SRN] [38]	MJ+SJ	word	95.5	91.5	94.8	82.7	85.1	87.8
Ours	DPAN	MJ+SJ	word	97.3	93.5	95.9	84.9	88.2	91.2
	DPAN-wiki	MJ+SJ+wiki	word	97.7	93.9	96.2	85.5	89.0	91.9

Figure 5: The visualization of G_1 (left column of a,b) and G_2 (right column of a,b). Visual misalignments exist in G_1 for some positions, which affects the recognition results.

profitable than pre-training one of them; (2) Only pre-training the BLM degrades the performance; (3) Although stacking transformer units improves the performance, there is an upper limit. To maximize the performance of DPAN, we take the model corresponding to the last row of Tab.3 as the ultimate model to be evaluated.

4.4 Comparison with State-Of-the-Art Methods

We compare our method with the current state-of-the-art methods in Tab.4 where only the results without a lexicon are listed. To further boost the accuracy and robustness of DPAN, we implement another model DPAN-wiki where the BLM is pre-trained on wikitext-103 [20]. As can be seen from the results shown in Tab.4, DPAN achieves the best performance among all types of approaches. Note that GTC[8] uses additional text images for training. Even so, DPAN performs better than GTC on most datasets. Compared to the PDED-based methods, DPAN has a significant lead in accuracy on each dataset. In particular, DPAN surpasses these two methods [19, 38] on SVT, SVTP, and CUTE80 by large margins as these datasets include a lot of blurred text images and irregular text images which limit the ability of PPAM. Moreover, when a larger amount of text data is used to pre-train BLM, the model DPAN-wiki gains accuracy promotions on IC13, SVT, IIIT 5K, SVT and CUTE80

Figure 6: The successfully recognized hard samples by DPAN. The blurry texts, curved texts and texts with abnormal fonts are accurately recognized.

by 0.4%, 0.4%, 0.6%, 0.8% and 0.7%, respectively. We attribute the performance improvements to the text diversity and better data distribution introduced by wikitext-103, which enables BLM to learn more robust linguistic contexts to facilitate the process of PCAM and decoding. Some qualitative illustrations are shown in Fig.6, reflecting the robustness of DPAN on hard samples

5 CONCLUSION

In this paper, we aim to alleviate the visual misalignment problem of parallel-decoupled encoder-decoder (PDED) framework on hard samples. To deal with this problem, we propose a dual parallel attention network (DPAN), where a parallel contextual attention (PCAM) is present to rectify the inaccurate visual glimpses generated from PPAM. The proposed PCAM leverages a bidirectional language model to enhance the output features from PPAM with linguistic contexts, and considers them as contextualized queries to do more accurate character region attention. The evaluation results on 6 benchmarks show that DPAN achieves the state-of-the-art performance, verifying its effectiveness and robustness.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (2018YFB0804203), the National Nature Science Foundation of China (62022076, U1936210, 62032006), the Youth Innovation Promotion Association Chinese Academy of Sciences (2017209), and the Fundamental Research Funds for the Central Universities under Grant (WK3480000011).

REFERENCES

- [1] Christian Bartz, Joseph Bethge, Haojin Yang, and Christoph Meinel. 2019. Kiss: Keeping it simple for scene text recognition. *arXiv preprint arXiv:1911.08400* (2019).
- [2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*. Springer, 213–229.
- [3] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. 2017. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE international conference on computer vision*. 5076–5084.
- [4] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*. 369–376.
- [5] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. 2016. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2315–2324.
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [7] Pan He, Weilin Huang, Yu Qiao, Chen Loy, and Xiaoou Tang. 2016. Reading scene text in deep convolutional sequences. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.
- [8] Wenyang Hu, Xiaocong Cai, Jun Hou, Shuai Yi, and Zhiping Lin. 2020. Gtc: Guided training of ctc towards efficient and accurate scene text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11005–11012.
- [9] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227* (2014).
- [10] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2016. Reading text in the wild with convolutional neural networks. *International journal of computer vision* 116, 1 (2016), 1–20.
- [11] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. 2015. Spatial transformer networks. *arXiv preprint arXiv:1506.02025* (2015).
- [12] Dimosthenis Karatzas, Lluís Gomez-Bigorda, Angelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. 2015. ICDAR 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1156–1160.
- [13] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluís Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluís Pere De Las Heras. 2013. ICDAR 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition*. IEEE, 1484–1493.
- [14] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. 2019. Show, attend and read: A simple and strong baseline for irregular text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8610–8617.
- [15] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. 2017. Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2359–2367.
- [16] Minghui Liao, Pengyuan Lyu, Minghang He, Cong Yao, Wenhao Wu, and Xiang Bai. 2019. Mask TextSpotter: An End-to-End Trainable Neural Network for Spotting Text with Arbitrary Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).
- [17] Minghui Liao, Jian Zhang, Zhaoyi Wan, Fengming Xie, Jiajun Liang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2019. Scene text recognition from two-dimensional perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8714–8721.
- [18] Wei Liu, Chaofeng Chen, Kwan-Yee K Wong, Zhizhong Su, and Junyu Han. 2016. STAR-Net: a spatial attention residue network for scene text recognition. In *BMVC*, Vol. 2. 7.
- [19] Pengyuan Lyu, Zhicheng Yang, Xinhang Leng, Xiaojun Wu, Ruiyu Li, and Xiaoyong Shen. 2019. 2d attentional irregular scene text recognizer. *arXiv preprint arXiv:1906.05708* (2019).
- [20] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843* (2016).
- [21] Anand Mishra, Karteek Alahari, and CV Jawahar. 2012. Scene text recognition using higher order language priors. In *BMVC-British Machine Vision Conference*. BMVA.
- [22] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. 2013. Recognizing text with perspective distortion in natural scenes. In *Proceedings of the IEEE International Conference on Computer Vision*. 569–576.
- [23] Zhi Qiao, Yu Zhou, Dongbao Yang, Yucan Zhou, and Weiping Wang. 2020. Seed: Semantics enhanced encoder-decoder framework for scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13528–13537.
- [24] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. 2014. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications* 41, 18 (2014), 8027–8048.
- [25] Baoguang Shi, Xiang Bai, and Cong Yao. 2016. 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* 39, 11 (2016), 2298–2304.
- [26] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2016. Robust scene text recognition with automatic rectification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4168–4176.
- [27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).
- [28] Zhaoyi Wan, Minghang He, Haoran Chen, Xiang Bai, and Cong Yao. 2020. TextScanner: Reading characters in order for robust scene text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12120–12127.
- [29] Zhaoyi Wan, Fengming Xie, Yibo Liu, Xiang Bai, and Cong Yao. 2019. 2d-ctc for scene text recognition. *arXiv preprint arXiv:1907.09705* (2019).
- [30] Kai Wang, Boris Babenko, and Serge Belongie. 2011. End-to-end scene text recognition. In *2011 International Conference on Computer Vision*. IEEE, 1457–1464.
- [31] Tianwei Wang, Yuanzhi Zhu, Lianwen Jin, Canjie Luo, Xiaoxue Chen, Yaqiang Wu, Qianying Wang, and Mingxiang Cai. 2020. Decoupled attention network for text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12216–12224.
- [32] Lin Wu, Yang Wang, Xue Li, and Junbin Gao. 2018. Deep attention-based spatially recursive networks for fine-grained visual recognition. *IEEE transactions on cybernetics* 49, 5 (2018), 1791–1802.
- [33] Lin Wu, Yang Wang, Hongzhi Yin, Meng Wang, and Ling Shao. 2019. Few-shot deep adversarial learning for video-based person re-identification. *IEEE Transactions on Image Processing* 29 (2019), 1233–1245.
- [34] Liang Wu, Chengquan Zhang, Jiaming Liu, Junyu Han, Jingtuo Liu, Errui Ding, and Xiang Bai. 2019. Editing Text in the Wild. In *Proceedings of the 27th ACM International Conference on Multimedia (Nice, France) (MM '19)*. Association for Computing Machinery, New York, NY, USA, 1500–1508. <https://doi.org/10.1145/3343031.3350929>
- [35] Rui Yan and Yaohong Huang. [n.d.]. PlugNet: Degradation Aware Scene Text Recognition Supervised by a Pluggable Super-Resolution Unit. ([n.d.]).
- [36] Mingkun Yang, Yushuo Guan, Minghui Liao, Xin He, Kaigui Bian, Song Bai, Cong Yao, and Xiang Bai. 2019. Symmetry-constrained rectification network for scene text recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9147–9156.
- [37] Xiao Yang, Dafang He, Zihan Zhou, Daniel Kifer, and C Lee Giles. 2017. Learning to Read Irregular Text with Attention Mechanisms. In *IJCAI*, Vol. 1. 3.
- [38] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. 2020. Towards accurate scene text recognition with semantic reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12113–12122.
- [39] Hongyuan Yu, Yan Huang, Lihong Pi, Chengquan Zhang, Xuan Li, and Liang Wang. 2021. End-to-end video text detection with online tracking. *Pattern Recognition* 113 (2021), 107791.
- [40] Xiaoyu Yue, Zhanghui Kuang, Chenhao Lin, Hongbin Sun, and Wayne Zhang. 2020. RobustScanner: Dynamically Enhancing Positional Clues for Robust Text Recognition. In *European Conference on Computer Vision*. Springer, 135–151.
- [41] Fangneng Zhan and Shijian Lu. 2019. Esir: End-to-end scene text recognition via iterative image rectification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2059–2068.