Looking from a Higher-level Perspective: Attention and Recognition Enhanced Multi-scale Scene Text Segmentation

Yujin Ren[1 [0000−0001−9647−6713]], Jiaxin Zhang[1 [0000−0001−9787−9514]], Bangdong Chen[1 [0000−0002−6624−6375]], Xiaoyi Zhang[1 [0000−0002−5345−2110]], and Lianwen Jin[1,2⋆ [0000−0002−5456−0957]]

[1] South China University of Technology, Guangzhou, China
[2] SCUT-Zhuhai Institute of Modern Industrial Innovation, Zhuhai, China
yujinren98@gmail.com,{msjxzhang,eebdchen}@mail.scut.edu.cn
xy_zhang@foxmail.com,lianwen.jin@gmail.com

Abstract. Scene text segmentation, which aims to generate pixel-level text masks, is an integral part of many fine-grained text tasks, such as text editing and text removal. Multi-scale irregular scene texts are often trapped in complex background noise around the image, and their textures are diverse and sometimes even similar to those of the background. These specific problems bring challenges that make general segmentation methods ineffective in the context of scene text. To tackle the aforementioned issues, we propose a new scene text segmentation pipeline called Attention and Recognition enhanced Multi-scale segmentation Network (ARM-Net), which consists of three main components: Text Segmentation Module (TSM) generates rectangular receptive fields of various sizes to fit scene text and integrate global information adequately; Dual Perceptual Decoder (DPD) strengthens the connection between pixels that belong to the same category from the spatial and channel perspective simultaneously during upsampling, and Recognition Enhanced Module (REM) provides text attention maps as a prior for the segmentation network, which can inherently distinguish text from background noise. Via extensive experiments, we demonstrate the effectiveness of each module of ARM-Net, and its performance surpasses that of existing state-of-the-art scene text segmentation methods. We also show that the pixel-level mask produced by our method can further improve the performance of text removal and scene text recognition.

Keywords: Scene Text Segmentation · Deep Neural Network.

1 Introduction

As an important constituent of image pre-processing, text segmentation was once the foundation of text detection and recognition. With mature applications

⋆ Corresponding author.

Fig. 1. Three specific issues in scene text segmentation [4]: (a) various text scales; (b) scattered text distribution; (c) background distraction.

of deep neural networks (DNNs) in optical character recognition (OCR), pixel-level (stroke) text segmentation is rarely used in traditional text-related vision tasks. However, some more fine-grained scenarios have emerged recently, such as text editing [39, 25, 14] and text removal [44, 17]. They require segmentation to obtain precise pixel-level text masks in advance, which can be used to separate texts from complex backgrounds. In response to these new demands, scene text segmentation has gradually regained researchers' attention [23, 1, 32, 40, 41].

It is difficult to obtain satisfactory results by directly transferring general segmentation methods to scene text, as there are specific issues that must be addressed in scene text segmentation: (1) Scene text is non-convex and prone to exhibiting drastic differences in scale, making it challenging to segment structural details of texts with various styles. (2) The uneven distribution of scene text in the image makes it easy for text that appears in inconspicuous locations, especially text whose texture appears less frequently, to be ignored by the segmentation network. (3) Scene text is trapped in complex background noise and sometimes has similar textures with them, which may lead to ambiguity in segmentation results.

Although existing scene text segmentation approaches partially solve the aforementioned problems to some extent, they still have major shortcomings. SMANet [1] used a pooling operation to obtain multi-scale text features, but it cannot preserve the resolution of the feature map, making it unsuitable for retaining the spatial location information of scene text. MGNet [32] adopted a semi-supervised training strategy and used polygon-level mask annotations to provide a prior for pixel-level text segmentation, which is helpful to confirm text location, but lacks a subtle network design for scene text. TexRNet [40] well-designed a refinement network after the segmentation backbone, exploiting cosine similarity to correct those infrequent pixels that are misclassified. It can indeed produce a better segmentation effect by considering the characteristics of scene text, but requires complex character-level annotations for its discriminator.

In this study, we propose a text-tailored segmentation pipeline called ARM-Net, which jointly focuses on both low-level appearance information and higher-level text semantic information. It is worth noting that low-level and high-level

features in segmentation network are collectively referred to as low-level text appearance information to differentiate them from text semantic information. We optimize low-level text appearance information by rethinking the classical encoder-decoder structure of the segmentation network. In the feature encoding stage, the proposed Text Segmentation Module (TSM) is used for modeling sophisticated text segmentation features by accommodating global and local perspectives. It assigns equal attention weight to global texts to reactivate those with rare textures because of their strong semantic association with the dominant text. Moreover, it also adapts irregular multi-scale scene text to eliminate the interference of background noise and thus capture more effective local features. In the decoding recovery stage, pixels are progressively aggregated into specific classes during upsampling. Slight deviations in deep feature maps may result in inaccurate and distorted segmentation, especially on scene text with an arbitrary shape. Accordingly, we propose a Dual Perceptual Decoder (DPD), whose parameters can be dynamically adjusted to spatial and channel contents. Aiming to take full advantage of text characteristics, we explore the essential differences between text and generic scenes (background noise in scene text segmentation), explaining why human beings rarely struggle with how to distinguish between them. The key, we believe, is that text is no longer treated as simple graphic symbols after people endow them with specific meanings. To imitate the human behavioral patterns, we design an innovative Recognition Enhanced Module (REM) to introduce higher-level text semantic information that provides text attention maps as prior knowledge to promote text discrimination.

To summarize, our main contributions are three-fold:

1. We propose an end-to-end trainable model, ARM-Net, which exploits a combination of low-level text appearance information and higher-level text semantic information to facilitate segmentation.
2. Extensive experiments demonstrate the effectiveness of ARM-Net, which achieves superior performance on three mainstream scene text segmentation benchmarks, and each module plays significant role.
3. Experiments on downstream tasks illustrate that the addition of pixel-level masks generated by ARM-Net can improve the effectiveness of text removal and the accuracy of scene text recognition.

2 Related Work

2.1 Semantic Segmentation

Semantic segmentation, one of the traditional tasks in computer vision, aims to predict a correct category for each pixel. With the development of deep learning, many methods based on the DNN are distinguishable from traditional graph algorithms such as MRF [31] and CRF [15]. Since the FCN [19] firstly adopted fully convolutional network in semantic segmentation, numerous works based on the encoder-decoder structure [24, 46, 8] have emerged.

In order to overcome the dilemma of limited receptive fields, the importance of multi-scale features has been continuously emphasized. PSPNet [45] fused

features of different scales through pooling operations to aggregate contextual information in different regions. DeepLab [2, 3] introduced the atrous convolution operation to obtain multi-scale features without changing the resolution of feature maps. HRNet [35] performed repetitive fusion by exchanging information on parallel multi-resolution sub-networks so that the network can maintain high-resolution representation.

As the self-attention mechanism [29] has shown extraordinary value in natural language processing (NLP), researchers have applied it to semantic segmentation to obtain the long-range dependency. DANet [5] associated spatial attention and channel attention to acquire broader contextual information. Employing a similar strategy, CCNet [9] and Axial-DeepLab [33] proposed Criss-Cross Attention and Axial-Attention, which both use fewer pixels to participate in the attention calculation so as to reduce computing costs. EMANet [16] abandoned the process of computing attention maps on a full graph and utilized the Expectation Maximization (EM) algorithm instead to iterate over a set of bases and then performed the attention mechanism on it.

2.2 Scene Text Segmentation

Previously developed scene text segmentation methods mostly use thresholds or low-level features to binarize scene text images, making it difficult to produce satisfactory results due to the complexity of scenes and textures. Recently, several approaches based on deep learning have been explored. Bonechi et al. [1] labeled COCO [30] and MLT [22] by machine for pre-training, and proposed SMANet, which combines the pooling pyramid with the attention mechanism to form a multi-scale attention module. Wang et al. [32] proposed a mutually guided dual-task network that uses polygon-level masks (bounding boxes) for semi-supervision which can be easily obtained from scene text detection datasets. They pointed out that pixels outside a polygon-level mask do not belong to the pixel-level mask. So, polygon-level masks can serve as a filter to guide the generation of pixel-level masks, and vice versa. Xu et al. [40] made a new text segmentation dataset, TextSeg, which contains 4,024 images with comprehensive annotations. They made some special designs in their TexRNet to refine the output from the aforementioned segmentation networks such as DeepLabV3+ [3] and HRNet [35] to improve their performance on scene text segmentation. TexRNet firstly guarantees that high-confidence regions are reliable by calculating the modified cosine-similarity between the text class and background class. It uses the key features pooling and attention-based similarity checking to activate text regions that may be ignored owing to low-confidence in the initial prediction.

3 Methodology

3.1 Pipeline

As shown in Fig. 2, our proposed ARM-Net consists of three main components: Text Segmentation Module (TSM), Dual Perceptual Decoder (DPD) and Recognition Enhanced Module (REM). We adopt ResNet-50 [6] as backbone for feature

extraction. Extracted features are fed into the TSM and transformed to dense multi-scale text segmentation features with a global view. Then, the DPD dynamically aggregates text and background pixels according to content information from the spatial and channel perspective during upsampling. We also blend low-level features from the backbone with each stage of the DPD to acquire more visual details. In addition, the REM is applied as an auxiliary cue and brings in higher-level text semantic information to enhance segmentation features. Text attention maps generated by the REM indicate where the segmentation network should focus.

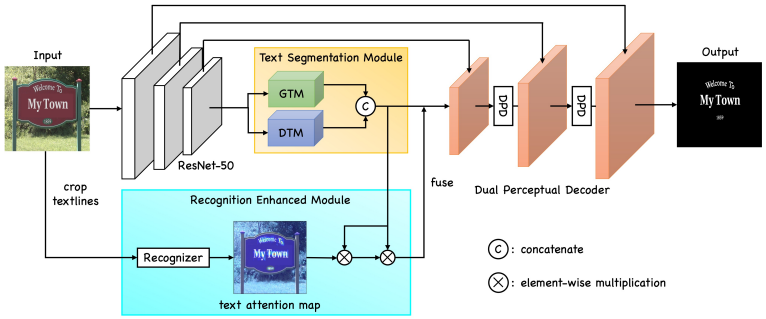

Fig. 2. Pipeline of our proposed ARM-Net.

3.2 Text Segmentation Module

For scene text segmentation, local and global information are like both sides of a scale, and a reasonable balance needs to be achieved between the two to succeed better performance. Accordingly, our proposed TSM integrates global correlations and local details adequately to obtain more effective segmentation representations, as illustrated in Fig. 3.

Scene texts are frequently scattered in images, and they are sometimes submerged in the complex background noise; so, those texts with small size or rare textures are easily ignored by the segmentation network. We propose a Global Text Module (GTM), which draws on the core idea of CCNet [9] to model dependencies of the entire image, while avoiding a surge of computation and parameters. Unlike the non-local network [38] that calculates the correlation matrix between each pixel in the feature map spatially, we only perform the self-attention for each pixel in the horizontal and vertical directions that the pixel belongs to, as shown in the red dashed box of Fig. 3.

In the concrete, after 1×1 convolution layers, we obtain three feature maps Q, K and V. The affinity matrix $A = \varphi\left(Q_u \cdot K_u^T\right)$ between spatial locations is obtained by element-wise multiplying each position of Q (i.e. Q_u) with the vector set $K_u \in \mathbb{R}^{C' \times (H+W-1)}$ and then applying a softmax function, $\varphi(\cdot)$. Here

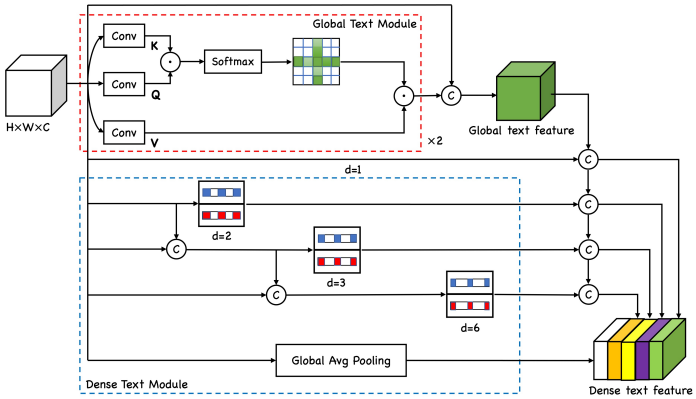

Fig. 3. Architecture of the TSM, which consists of GTM and DTM two sub-modules. In the DTM, blue and red squares represent horizontal and vertical atrous convolution, respectively, and d is the dilation rate.

K_u denotes the set of positions in K that are in the same row or column as u. Similar to the above operation, $V_u \in \mathbb{R}^{C\times(H+W-1)}$ is multiplied with the affinity matrix A. Then, we add it to the primary local feature, H, to obtain the pixel-wise contextual augmented feature representation, $H^{'}$. Here we implement Criss-Cross Attention twice to deliver the information of one pixel into all paths. The above operation can be expressed as follows:

$$H_u^{'} = \varphi\left(Q_u \cdot K_u^T\right) \cdot V_u + H_u \ . \tag{1}$$

As scene texts are mostly in rectangular or curved shape and their scale is extremely different, we propose a Dense Text Module (DTM) on the basis of atrous convolution [2]. The operation of atrous convolution is equivalent to inserting $d-1$ zeros between two adjacent weights of the filter, where d is the dilation rate. This approach can expand the range of the receptive field while maintaining the resolution of the feature map.

As illustrated in the blue dashed box of Fig. 3, the DTM heuristically performs atrous convolution in two directions, which separately creates horizontal or vertical zero padding. Afterwards, the DTM cascades these atrous convolution layers with the dilation rate from low to high. In this way, we obtain denser feature representations and rectangular receptive fields with various aspect ratio, which are more appropriate for irregular scene text. When the dilation rates of the two directions are equal, the receptive field is of a regular shape. Stacking n atrous convolution layers can obtain a larger equivalent kernel size K:

$$K = \sum_{i=1}^{n} K_i - (n-1) \ . \tag{2}$$

Moreover, the receptive field size, R, increases linearly with dilation rates:

$$R = (d-1) \times (K-1) + K \ . \tag{3}$$

Following equations (2) and (3), the number of R in the DTM is at most 26, which sufficiently alleviates the problem wherein the fixed shape receptive field is not appropriate for multi-scale scene text.

The DTM outputs dense multi-scale text features with abundant sizes of the receptive field. Each feature is concatenated with the augmented feature representation $H^{'}$ produced by the GTM to combine global information adequately.

3.3 Dual Perceptual Decoder

In segmentation methods that utilize the encoder-decoder structure, decoder plays an important role in rebuilding image from features. However, traditional upsampling methods have their limitations. Nearest and bilinear interpolation only consider adjacent positions, and deconvolution is restrained by fixed kernel size and weights. To better cope with the problem that are neglected by previous scene text segmentation methods, we propose the DPD, as shown in Fig. 4.

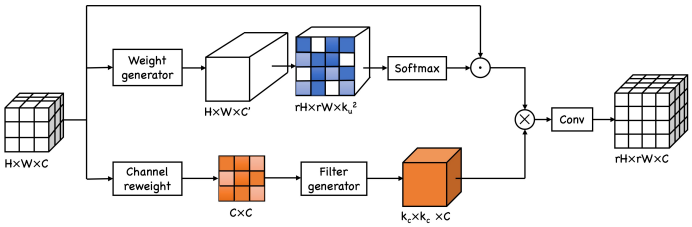

Fig. 4. Structure of the DPD. The upper branch is Spatial Context-Aware Module, and the branch below is Channel Semantic-Aware Module.

The upper branch in Fig. 4 is a Spatial Context-Aware Module, which is inspired by CARAFE [34] and the dynamic filter network [11]. Given a $C \times H \times W$ feature map, F, the Spatial Context-Aware Module can dynamically expand the feature map to $C \times rH \times rW$ according to the context information of different objects in the spatial dimension, where r is the up-sampling rate. We first employ a convolution layer with kernel size k_e on segmentation features as a weight generator through which the feature map becomes $H \times W \times C'$. Here, $C' = r^2 k_u^2$, and k_u is the kernel size during up-sampling. The weight generator aggregates the spatial context information within the $k_e \times k_e$ receptive field. Then, we reshape the channel dimension and spatial dimension to obtain a $k_u \times k_u$ weight matrix, $W_u \in \mathbb{R}^{rH \times rW \times k_u^2}$, which can satisfy each position on $C \times rH \times rW$ as an individual weight. Note that each $k_u \times k_u$ kernel is normalized by a spatial softmax function. This makes the kernel values sum to one and has no effect on the feature distribution. Finally, we calculate the upsampled output, F_s, as follows:

$$F_s = \sum_{i=-n}^{n} W_i \cdot F_i \ . \tag{4}$$

From another perspective, each channel of high-level features can be regarded as a class-specific response, and semantic responses belonging to the same category should be associated with each other during upsampling. By assigning channel weights according to their interdependencies, associated channels are emphasized, and interfering channels are suppressed simultaneously, thereby improving the discriminative capacity of the model.

To fulfill the above purpose, we design the Channel Semantic-Aware Module as illustrated in the lower branch of Fig. 4. Taking segmentation feature $F \in \mathbb{R}^{C\times H\times W}$ as the input, we first calculate its channel-wise relationship on it. Concretely, we reshape F to $\mathbb{R}^{C\times N}$, and then multiply it with its transpose matrix. We normalize the output through a softmax layer to obtain the weight matrix $W_c \in \mathbb{R}^{C\times C}$. W_c represents the interrelationship between channels and can also be seen as a channel attention map.

$$w_c = \frac{exp(F_i \cdot F_i^T)}{\sum_{i=1}^{C} exp(F_i \cdot F_i^T)} , \tag{5}$$

where w_c is an element of W_c that measures the degree of correlation between two channels. Sequentially, a filter generation layer, g, is applied to generate channel semantic-aware features. Unlike SENet [7], which squeezes the spatial information of each feature map into one channel descriptor by global average pooling, we use adaptive average pooling to generate a $k \times k$ channel-wise feature F_c. Afterward, F_c can be viewed as a channel-weighted dynamic filter, and each position in the filter represents general information of a sub-region.

Finally, a depth-wise convolution layer with F_c as weights is applied to F_s. Through a 1×1 convolution layer to fuse the channel information, upsampled output is obtained. Replacing traditional upsampling methods with the DPD allows us to rearrange activation responses of scene text elaborately according to the spatial and channel information.

3.4 Recognition Enhanced Module

When the texture is similar to the background, scene text is difficult to be identified accurately by the segmentation network. We think it is because the segmentation network only utilizes low-level appearance information such as text structure and color. To alleviate this problem, the network needs to be taught that text has higher-level meanings that go beyond general objects and symbols, just like human beings do. Consequently, we propose an REM, which can highlight text regions in the whole image and deliver higher-level text semantic information to segmentation features, as shown in Fig. 5.

In the training phase, we first crop textlines from the image and feed them into a pre-trained DAN [37] recognizer whose parameters are frozen. The reason why we choose DAN is that it mitigates attention drift problem through the Convolutional Alignment Module (CAM). Other attention-based recognizers [21, 27] are also appropriate here. The cropped textline images are unified into $M \times 128 \times 32$, where M is the total number of textlines in one batch. Attention maps produced by DAN are used to align character positions at each time

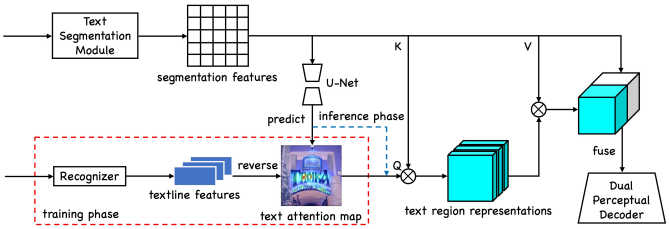

Fig. 5. Architecture of the REM. Portions in the red dotted box only participate in the training phase, and during inference, the text attention map is predicted by U-Net.

step, which are implicitly supervised by word annotations. Summing all steps' attention maps along the channel can obtain a textline attention map. Furthermore, we put textline attention maps into an all-black background according to their initial positions to acquire the text attention map of the whole image, which can indicate text distribution. Pixels with a higher probability in the text attention map are more likely to be candidates of text regions. For a better relation estimation, we follow the operation in the self-attention mechanism [29]. Formally, we treat the text attention map as Q and segmentation features as K, V. After computing the relation matrix between the text attention map, F_Q, and segmentation features, F_K, we assign it as text region representations to the corresponding position of segmentation features, F_V. As a result, the segmentation feature with high similarity to the corresponding position on the text attention map is highlighted while that with low similarity is restrained. By this means, we obtain the enhanced text features, T_e, as follows:

$$T_e = \frac{exp(F_K \cdot F_Q^T)}{\sum_{i=1}^{n} exp(F_K \cdot F_Q^T)} \cdot F_V \ . \tag{6}$$

Since the coordinate information of texts is not available in the inference phase, we employ a lightweight U-Net [24] to predict text attention maps by using segmentation features under the supervision of the l_1 loss. In this way, we can not only impose a more direct constraint on segmentation features but also replace text attention maps' output by recognizer with that prediction, enabling an image-to-image inference process.

4 Experiment

4.1 Datasets and Implementation Details

We conduct experiments on the following three benchmarks, all of which have high-quality pixel-level annotations that can be used to supervise the training phase of segmentation network.

1) **ICDAR-2013** [12]: This is a dataset for the ICDAR 2013 'Robust Text Reading' competition, which contains 229 training images and 233 test images

with pixel-level mask ground-truth. Scene texts in this dataset are regular and can be surrounded by rectangular bounding box.

2) **Total-Text** [4]: It contains 1,255 training images and 300 test images. Unlike ICDAR-2013, scene texts in Total-Text have irregular shapes, including rectangular and curved texts. Clear pixel-level annotations are also available.

3) **TextSeg** [40]: TextSeg consists of 4,024 scene text and design text images, which are split into training, validation, and testing sets, with 2,646, 340, and 1,038 images, respectively. TextSeg provides accurate and comprehensive annotations including word-wise and character-wise polygon-level masks as well as pixel-level masks and transcriptions.

We train each model on a NVIDA V100 GPU for 100 epochs and use the Adam [13] optimizer. The first five epochs are warm-ups, and the remaining epochs employ poly decayed learning rates, where the initial learning rate 1e-4 is multiplied by $\left(1-\frac{iter}{max_{iter}}\right)^{power}$ with power=0.9.

4.2 Comparison with Existing Methods

We compare our proposed ARM-Net with some state-of-the-art methods, including three semantic segmentation methods [24, 3, 35] and three text segmentation methods [1, 32, 40]. We use the Precision(P), Recall(R), and F-score(F) of the foreground text to quantitatively evaluate the performance of the network, where $F=\frac{2P\cdot R}{P+R}$ denotes the harmonic mean of P and R. The results of three evaluation metrics are presented in Table 1. ARM-Net outperforms the state-of-the-art methods in pixel-level scene text segmentation on all three benchmarks. Furthermore, the inference speed of ARM-Net (4.35fps) is faster than TexRNet (1.22 fps), which shows that our method has a better efficiency.

Table 1. Quantitative results between ARM-Net and other segmentation methods.

Methods	Total-Text			ICDAR 2013			TextSeg		
	P	R	F	P	R	F	P	R	F
U-Net [24] (MICCAI'15)	79.6	69.7	74.3	74.6	53.9	62.6	89.0	77.4	82.8
DeepLabV3+ [3] (ECCV'18)	80.2	76.5	78.3	77.4	63.2	69.6	91.4	90.9	91.2
HRNetV2 [35] (TPAMI'20)	81.4	78.0	79.7	72.8	69.5	71.1	91.9	90.5	91.2
SMANet [1] (PRL'20)	86.6	73.9	77.5	74.4	73.8	71.3	-	-	-
MGNet [32] (TIP'20)	83.3	81.6	80.5	79.0	77.0	74.5	-	-	-
TexRNet+DeepLabV3+ [40] (CVPR'21)	-	-	84.4	-	-	83.5	-	-	92.1
TexRNet+HRNetV2-W48 [40] (CVPR'21)	-	-	84.8	-	-	85.0	-	-	92.4
ARM-Net (Ours)	**87.1**	**83.8**	**85.4**	**88.9**	**81.6**	**85.1**	**92.8**	**92.6**	**92.7**

Several typical qualitative examples are presented in Fig. 6, where images contain a variety of styles/shapes/arrangements of text, as well as other complex background distractions such as illumination. It is obvious that our approach overcomes these difficulties and produces more accurate results.

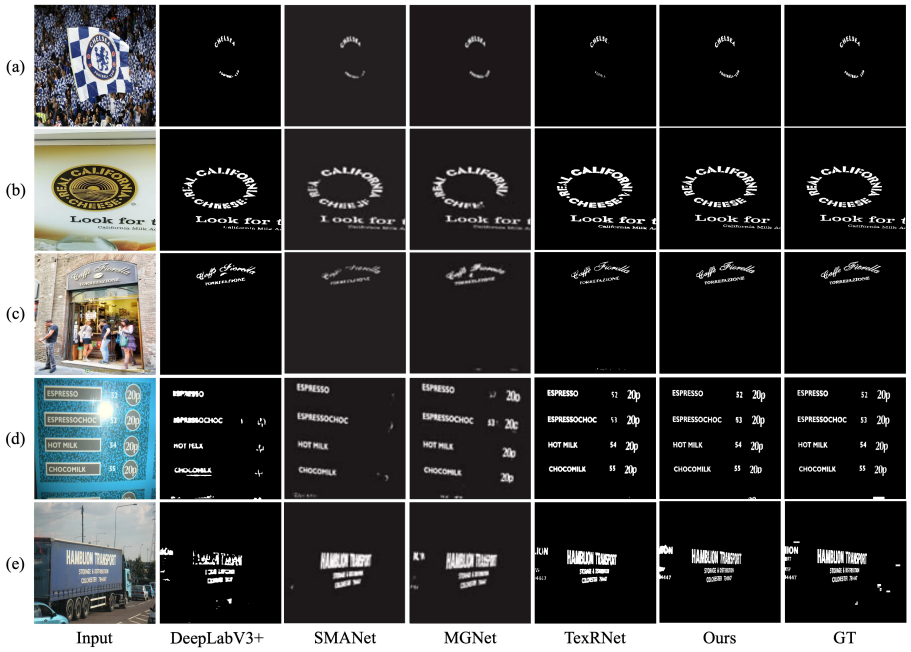

Fig. 6. Text segmentation visualization results on Total-Text(a∼c) and ICDAR-2013(d,e). From left to right, each column is input, masks predicted by DeepLabV3+, SMANet, MGNet, TexRNet, our ARM-Net, and ground truth, respectively.

4.3 Ablation Studies

In this section, we conduct ablation experiments to verify the effectiveness of each module in our method. All ablation experiments are performed on Total-Text. Note that when conducting ablation experiments within one module, we ensure that other modules are in the best settings.

We first investigate three main components of ARM-Net. The baseline is a simple FCN [19], beginning with which, we add the TSM after the encoder, replace the decoder with the DPD, and introduce the REM progressively. The results in Table 2 exhibit a continuous upward trend with the introduction of each module. When including TSM and DPD, the F-score increases 2.7%, while this gain is 2.5% when we use REM only. Moreover, the ARM-Net with all three modules achieves the best performance, with an increase in F-score of more than 3% compared to the baseline. This suggest that higher-level semantic information is an effective supplement and is as critical as low-level appearance information.

In addition, the samples in Fig. 7 provide more intuitive evidence that ARM-Net is highly adaptable to the scene text segmentation task. Because of the comprehensive consideration of low-level text appearance information and higher-level text semantic information through TSM, DPD, and REM, our approach

Table 2. Effectiveness experiment on three main modules of ARM-Net.

TSM	DPD	REM	P	R	F
×	×	×	85.4	79.4	82.3
✓	×	×	85.7	83.4	84.6
✓	✓	×	85.9	**84.2**	85.0
×	×	✓	85.8	83.7	84.8
✓	✓	✓	**87.1**	83.8	**85.4**

achieves satisfactory segmentation results, without many serious cases of misclassification such as incomplete, blurred text structure and unfiltered background.

Fig. 7. Segmentation results for typical difficult samples. From top to bottom, each row is large variations in the scale and shape of text, scattered distribution of text throughout the image, and small text hidden in complex background noise, respectively.

Better low-level text appearance information. We conduct experiments in Table 3 to investigate the effectiveness of GTM and DTM. It can be seen that in scene text segmentation, both global attention information and local multi-scale information are beneficial for scene text segmentation. The shape and size of scene text are different from those of general objects; so, the dilation rate of DCM needs to be carefully designed to prevent degradation problem, that is, as the dilation rate increases, fewer weights are applied to valid feature regions. Consequently, we empirically set both the horizontal and vertical dilation rate to 1,2,3,6 based on the experimental results in Table 4.

Higher-level text semantic information. REM aims to deliver higher-level semantic information from the recognizer to the segmentation network. As demonstrated in Table 2: the addition of REM increases the F-score by a further 0.4%

Table 3. Comparison on the impact of GTM and DTM to text segmentation.

GTM	DTM	P	R	F
×	×	85.9	80.1	82.9
✓	×	87.0	80.5	83.6
×	✓	**87.2**	82.3	84.7
✓	✓	87.1	**83.8**	**85.4**

Table 4. Ablation experiment on dilation rate combinations of DTM.

Dilation rate	P	R	F
1,12,18,24	86.3	81.6	83.9
1,6,12,18	85.6	82.7	84.1
1,3,6,12	86.0	83.3	84.6
1,2,3,6	**87.1**	**83.8**	**85.4**

thus achieving optimal performance. As shown in Fig. 8, the role of REM is manifested in three aspects: (1) filtering out the misclassified background noise; (2) reactivating text regions that have been ignored by segmentation network; (3) distinguishing negative samples whose texture is close to that of the target text, eg. dot, symbols (e.g. '&'), and Chinese characters outside the ground truth. Characters that are not of concern to the recognizer are suppressed by REM along with the background, which is a similar pattern as that of humans, who treat unrecognized words as graphic symbols.

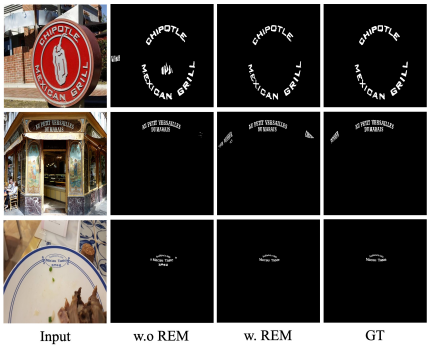

Fig. 8. Visualized results of using the REM (w. REM) and not (w.o REM).

Fig. 9. Downstream application on text removal task.

4.4 Downstream tasks

Downstream tasks such as text editing and text removal can benefit from fine-grained pixel-level text masks. Here we take text removal task as an example to demonstrate the application value of our ARM-Net. We feed the segmentation result as mask into DeepfillV2 [42], one of the state-of-the-art inpainting networks, to generate a text-free image. As shown in Fig. 9, the inpainting image using mask predicted by our method avoid suffering from serious smudging and erasing mistakes, which appear in inpainting reults using polygon masks from ground truth, and TexRNet (column 2 and 3). As DeepfillV2 utilizes the segmentation mask as a guidance, any misclassified pixels of text will be omitted

and those of background will be erased incorrectly. Such text removal results also reflects the superior segmentation performance of ARM-Net.

Apart from low-level image tasks, we observe that segmentation results also boost the performance of text recognition. Here we choose CRNN [26], a widely used method for text recognition, as the baseline. To utilize segmentation results, we simply concatenate pixel-level text masks and original images along the channel dimension, then feed it into the CRNN, which is initialized with an official pre-trained model, and fine-tune the first layer further with other parameters fixed. For a fair comparison, we train the model on the synthetic text dataset Synth90k [10] and validate it on ICDAR-2003 [20], ICDAR-2013 [12] and SVT [36], follow the setting of [26, 32]. The experimental results in Table 5 illustrate that the inclusion of pixel-level text masks generated by ARM-Net leads to an improvement in recognition accuracy of over 2% on both ICDAR-2003 and SVT, and of 1.2% on ICDAR-2013. The performance not only exceeds that of the CRNN baseline but is also better than the MGNet [32] method.

Table 5. Downstream experiment on scene text recognition.

Method	ICDAR-2003	ICDAR-2013	SVT
CRNN [26]	89.4	86.7	80.8
CRNN+MGNet [32]	91.4	87.7	82.8
CRNN+ARM-Net (Ours)	**91.7**	**87.9**	**82.9**

5 Conclusion

Scene text varies considerably in scale and shape, with some textures appearing infrequently, or even close to backgrounds. In this paper, we rethink the essence of scene text segmentation task and propose an effective end-to-end neural network, ARM-Net. The proposed TSM and DPD capture better low-level text appearance information, while the REM incorporates higher-level text semantic information as a complement. And by jointly exploiting both, we implement an optimization for the segmentation network. Quantitative and qualitative experiments demonstrate that our model outperforms state-of-the-art segmentation networks. We also show promising results that using text segmentation masks from ARM-Net on text removal and text recognition downstream tasks. In the future, we will investigate end-to-end networks for segmentation and recognition to further improve the performance of text segmentation in extreme scenarios.

Acknowledgements This research is supported in part by NSFC (Grant No.: 61936003), GD-NSF (no.2017A030312006, No.2021A1515011870), Zhuhai Industry Core and Key Technology Research Project (no. ZH22044702200058PJL), and the Science and Technology Foundation of Guangzhou Huangpu Development District (Grant 2020GH17)

References

1. Bonechi, S., Bianchini, M., Scarselli, F., Andreini, P.: Weak supervision for generating pixel–level annotations in scene text segmentation. Pattern Recognition Letters **138**, 1–7 (2020)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(4), 834–848 (2017)
3. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. pp. 801–818 (2018)
4. Ch'ng, C.K., Chan, C.S.: Total-text: A comprehensive dataset for scene text detection and recognition. In: 2017 14th IAPR International Conference on Document Analysis and Recognition. vol. 1, pp. 935–942. IEEE (2017)
5. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3146–3154 (2019)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141 (2018)
8. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: UNet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1055–1059. IEEE (2020)
9. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: CCNet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 603–612 (2019)
10. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and artificial neural networks for natural scene text recognition. arXiv preprint arXiv:1406.2227 (2014)
11. Jia, X., De Brabandere, B., Tuytelaars, T., Gool, L.V.: Dynamic filter networks. Advances in Neural Information Processing Systems **29** (2016)
12. Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L.G., Mestre, S.R., Mas, J., Mota, D.F., Almazan, J.A., De Las Heras, L.P.: ICDAR 2013 robust reading competition. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 1484–1493. IEEE (2013)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Krishnan, P., Kovvuri, R., Pang, G., Vassilev, B., Hassner, T.: TextStyleBrush: Transfer of text aesthetics from a single example. arXiv preprint arXiv:2106.08385 (2021)
15. Lafferty, J., Mccallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. proceedings of icml (2002)
16. Li, X., Zhong, Z., Wu, J., Yang, Y., Lin, Z., Liu, H.: Expectation-maximization attention networks for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9167–9176 (2019)

17. Liu, C., Liu, Y., Jin, L., Zhang, S., Luo, C., Wang, Y.: EraseNet: End-to-end text removal in the wild. IEEE Transactions on Image Processing **29**, 8760–8775 (2020)
18. Liu, R., Lehman, J., Molino, P., Petroski Such, F., Frank, E., Sergeev, A., Yosinski, J.: An intriguing failing of convolutional neural networks and the coordconv solution. Advances in Neural Information Processing Systems **31** (2018)
19. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440 (2015)
20. Lucas, S.M., Panaretos, A., Sosa, L., Tang, A., Wong, S., Young, R., Ashida, K., Nagai, H., Okamoto, M., Yamamoto, H., et al.: ICDAR 2003 robust reading competitions: entries, results, and future directions. International Journal of Document Analysis and Recognition **7**(2), 105–122 (2005)
21. Luo, C., Jin, L., Sun, Z.: Moran: A multi-object rectified attention network for scene text recognition. Pattern Recognition **90**, 109–118 (2019)
22. Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., Luo, Z., Pal, U., Rigaud, C., Chazalon, J., et al.: ICDAR 2017 robust reading challenge on multilingual scene text detection and script identification-rrc-mlt. In: 2017 14th IAPR International Conference on Document Analysis and Recognition. vol. 1, pp. 1454–1459. IEEE (2017)
23. Rong, X., Yi, C., Tian, Y.: Unambiguous scene text segmentation with referring expression comprehension. IEEE Transactions on Image Processing **29**, 591–601 (2019)
24. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention. pp. 234–241. Springer (2015)
25. Roy, P., Bhattacharya, S., Ghosh, S., Pal, U.: STEFANN: scene text editor using font adaptive neural network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13228–13237 (2020)
26. Shi, B., Bai, X., Yao, C.: 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), 2298–2304 (2016)
27. Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: Aster: An attentional scene text recognizer with flexible rectification. IEEE transactions on pattern analysis and machine intelligence **41**(9), 2035–2048 (2018)
28. Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J.: High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514 (2019)
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
30. Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S.: Coco-text: Dataset and benchmark for text detection and recognition in natural images. arXiv preprint arXiv:1601.07140 (2016)
31. Wang, C., Komodakis, N., Paragios, N.: Markov random field modeling, inference & learning in computer vision & image understanding: A survey. Computer Vision and Image Understanding **117**(11), 1610–1627 (2013)
32. Wang, C., Zhao, S., Zhu, L., Luo, K., Guo, Y., Wang, J., Liu, S.: Semi-supervised pixel-level scene text segmentation by mutually guided network. IEEE Transactions on Image Processing **30**, 8212–8221 (2021)

33. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.C.: Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In: European Conference on Computer Vision. pp. 108–126. Springer (2020)
34. Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C.C., Lin, D.: CARAFE: Content-aware reassembly of features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3007–3016 (2019)
35. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(10), 3349–3364 (2020)
36. Wang, K., Babenko, B., Belongie, S.: End-to-end scene text recognition. In: 2011 International Conference on Computer Cision. pp. 1457–1464. IEEE (2011)
37. Wang, T., Zhu, Y., Jin, L., Luo, C., Chen, X., Wu, Y., Wang, Q., Cai, M.: Decoupled attention network for text recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12216–12224 (2020)
38. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794–7803 (2018)
39. Wu, L., Zhang, C., Liu, J., Han, J., Liu, J., Ding, E., Bai, X.: Editing text in the wild. In: Proceedings of the 27th ACM international conference on multimedia. pp. 1500–1508 (2019)
40. Xu, X., Zhang, Z., Wang, Z., Price, B., Wang, Z., Shi, H.: Rethinking text segmentation: A novel dataset and a text-specific refinement approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12045–12055 (2021)
41. Xu, X., Qi, Z., Ma, J., Zhang, H., Shan, Y., Qie, X.: BTS: A bi-lingual benchmark for text segmentation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19152–19162 (2022)
42. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4471–4480 (2019)
43. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: European Conference on Computer Vision. pp. 173–190. Springer (2020)
44. Zhang, S., Liu, Y., Jin, L., Huang, Y., Lai, S.: Ensnet: Ensconce text in the wild. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 801–808 (2019)
45. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2881–2890 (2017)
46. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging **39**(6), 1856–1867 (2019)