Parallel Scale-wise Attention Network for Effective Scene Text Recognition

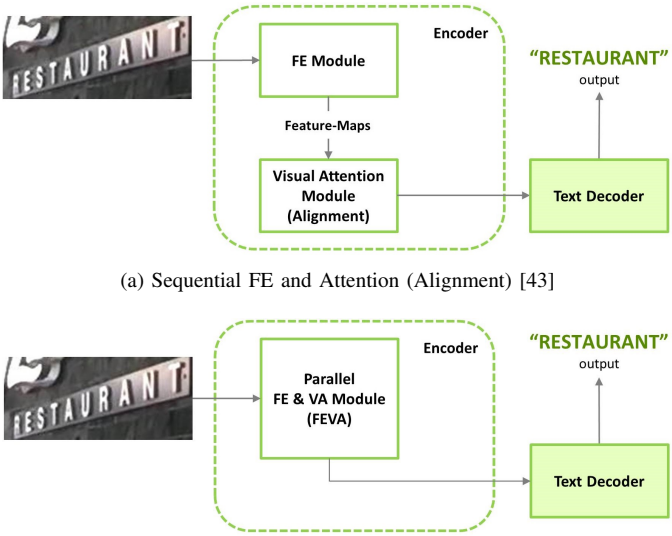
Usman Sajid[1], Michael Chow[2], Jin Zhang[2], Taejoon Kim[1], Guanghui Wang[3]
[1]*Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA, 66045*
[2]*Sony Interactive Entertainment Global R&D*
[3] *Department of Computer Science, Ryerson University, Toronto, ON, Canada M5B 2K3*
Email: {usajid, taejoonkim}@ku.edu[1], {michael.chow, jin.1.zhang}@sony.com[2], wangcs@ryerson.ca[3]

***Abstract*—The paper proposes a new text recognition network for scene-text images. Many state-of-the-art methods employ the attention mechanism either in the text encoder or decoder for the text alignment. Although the encoder-based attention yields promising results, these schemes inherit noticeable limitations. They perform the feature extraction (FE) and visual attention (VA) sequentially, which bounds the attention mechanism to rely only on the FE final single-scale output. Moreover, the utilization of the attention process is limited by only applying it directly to the single scale feature-maps. To address these issues, we propose a new multi-scale and encoder-based attention network for text recognition that performs the multi-scale FE and VA in parallel. The multi-scale channels also undergo regular fusion with each other to develop the coordinated knowledge together. Quantitative evaluation and robustness analysis on the standard benchmarks demonstrate that the proposed network outperforms the state-of-the-art in most cases.**

I. Introduction

Scene text recognition aims at extracting the screen text from the given input image. It serves as a trendy task in the computer vision field. The recognition task comes up with many key challenges and issues like huge background variation in and across different images, different font styles, big fluctuation in text appearance and scale. Automated text recognition remains more desirable as manual intervention proves to be very tedious and time-consuming. Recently, deep learning-based automated methods have shown superior performance in this domain and other tasks [18], [26], [31]–[33], [43], [47]. Some schemes perform character-level text recognition, while most methods do word/sentence level recognition. The latter one is more preferred due to its relatively easier and less tedious annotation process.

Among the best state-of-the-art deep networks, most of them [3], [6], [18], [34], [36], [43], [47] are based on the attention mechanism [2], [39]. The purpose of the attention mechanism is to align the text characters followed by their recognition. Generally, these methods incorporate the attention-based alignment and recognition into the decoder part of the network. However, these networks inherit an important limitation as the decoder gets highly over-burdened and sensitized with the dual task of text alignment and recognition. Consequently, it generates huge error propagation and aggregation within the decoder and thus compromises the effectiveness of the whole network. One possible solution is to decouple the attention/alignment mechanism from the decoder and integrate it with the feature extraction process inside the encoder block of the network. Recently, Wang *et al.* [43] proposed such decoupled attention network (DAN) with promising results. However, the encoder first sequentially performs the feature extraction (FE) followed by the visual attention (VA) process as shown in Fig. 1(a). This limits the DAN network efficacy as the attention mechanism only depends on and utilizes the final output feature-maps from the FE module. Consequently, the attention is not applied directly to each of the multi-scale feature-maps separately, but only to the final set of accumulated single-scale channels. Therefore, our focus revolves around the two main objectives in this work:



(a) Sequential FE and Attention (Alignment) [43]

(b) Parallel scale-wise FE and Alignment

Fig. 1. (a) The state-of-the-art Encoder-based attention mechanism [43] sequentially performs the feature extraction (FE) followed by the (attention) alignment module. (b) The proposed approach performs parallel FE and visual attention (VA) process on feature-maps with different scales within the encoder.

- Design a scale-wise visual attention-based scene text recognition network to address the key issues and challenges in this domain.
- Utilize the encoder-based and scale-wise attention

process in parallel to the feature extraction (FE) instead of standard sequential processing from FE to the visual attention module.

In this work, we propose a new multi-scale and scale-wise visually attended text recognition network to achieve the above objectives. As shown in Fig. 1(b), the feature extraction and visual alignment/attention (FEVA) have been done in parallel on different scale features within a single module, followed by the recognition-focused decoder to extract the scene text. In this way, we separately attend feature-maps from different scales directly instead of just attending the final single-scale channels. Moreover, we also deploy different and simpler visual attention process in contrast to the conventional deep up- and down-scaling fully connected networks (FCN) [21] based visual attention being used in DAN [43]. Several experiments on different standard benchmark datasets demonstrate the effectiveness of our scheme on both regular and irregular scene-texts as presented in the experiments section IV. The main contributions of this work include:

- We propose a new parallel FEVA-based encoder and multi-scale text recognition network to address the key recognition challenges and limitations in similar state-of-the-art architectures.
- We deploy the visual attention mechanism in an effective and unique way on multiple scales to enable the network in making a clearer distinction between the foreground and background pixels.
- Experimental evaluation on the standard benchmark datasets demonstrates that the proposed network outperforms the state-of-the-arts in most cases on both regular and irregular scene-texts.

II. Related Work

Text recognition problem remains a trendy topic in the computer vision field due to different challenges like varying text scale and size, partial occlusion, and non-axis aligned text. Before the deep learning era, document text recognition remained the main focus. [5] adopted the binarization process to extract the segmented text characters. But these methods are not applicable to scene-text due to different nature of issues like varying scale and style, and complex background. Most of the classical recognizers utilized the low-level information including the connected components [28], gradients descriptors (HoG) based on some feature-extraction mechanism [41]. Recently, deep-learning based method hugely surpass and outperform the traditional methods. They are categorized as segmentation-relying and segmentation-less text recognizers.

Segmentation-based methods undergo character-wise detection followed by the word formation. [4] designed five hidden fully-connected layers and ReLU Units [27] with softmax-based classification. [42] developed a convolutional neural network (CNN) with convolution and average pooling layers and the non-maximum suppression for character-wise text recognition. [14] used weight-shared CNN for three sub-tasks of dictionary, character sequence, and bag-of-N-gram encoding to perform the text recognition.

The segmentation-less schemes directly recognize the whole word or sentence from the given input image. [15] performed a CNN-based 90,000-way classification, where each category/class corresponds to one whole word. Shi *et al.* [34] integrated the convolutional neural network (CNN) and recurrent neural network (RNN)-based scheme to obtain the string features, and the Connectionist Temporal Classification (CTC)-based decoder to finally yield the recognized text. [35] employed the attention mechanism for text alignment before the recognition. Most following methods utilize the attention mechanism [2], [39] in one way or the other. Cheng *et al.* [6] designed the deep focused attention network (FAN) after observing and aiming to address the "attention-drift" problem in the recognition process, but it requires character-level annotations. [23], [35], [49] aimed at addressing non-axis aligned and distorted text via an attention-based mechanism. [1], [37] utilized the RNNs with Long Short Term Memory (LSTM) networks to perform the sequential word recognition. [11] integrated the CNN and RNN to design the deep-text recurrent network (DTRN) for recognizing text. Shi *et al.* [36] explicitly handled the text rectification by using the control points based rectification module and also applied the attention-based bi-directional LSTM decoder for text prediction. Li *et al.* [18] proposed a simple LSTM-based encoder-decoder framework via the 2D attention process. Wang *et al.* [43] designed the decoupled attention network (DAN) that performed the text alignment via convolution-based visual attention. Yu *et al.* [47] proposed the semantic reasoning network (SRN) for irregular scene-text that fuses the visual attention and semantic context modules while avoiding the RNN-based sequential processing.

Although these schemes produce good results, yet they fail to utilize the promising and beneficial attention mechanism explicitly on different multi-scale features. In this work, we work towards utilizing the multi-scale feature-extraction and visual-attention in parallel for better efficacy.

III. Proposed Approach

The paper proposes a new scene-text recognition network to address the major recognition challenges as detailed in Sec. I, as well as performs the visual attention explicitly on multi-scale feature-maps in parallel. The proposed network, as shown in Fig. 2, downscales the input image I ($\in \mathbb{R}^{3*H*W}$) resolution by half ($C * \frac{H}{2} * \frac{W}{2}$) using the initial convolutional layer. Here ($C = Channels, H = Height, W = Width$). The resultant feature-maps go through the text encoder (EN). The EN block comprises three parallel multi-scale modules (S1, S2, S3) with each module handling one specific scale. Each multi-scale module also visually attends the feature-maps in parallel to their conventional deep layers-based processing, followed by the concatenation together to generate their respective output. The visual attention helps the model to have a clearer understanding of the foreground and background pixels. Inspired by the high-resolution networks [38], [40],

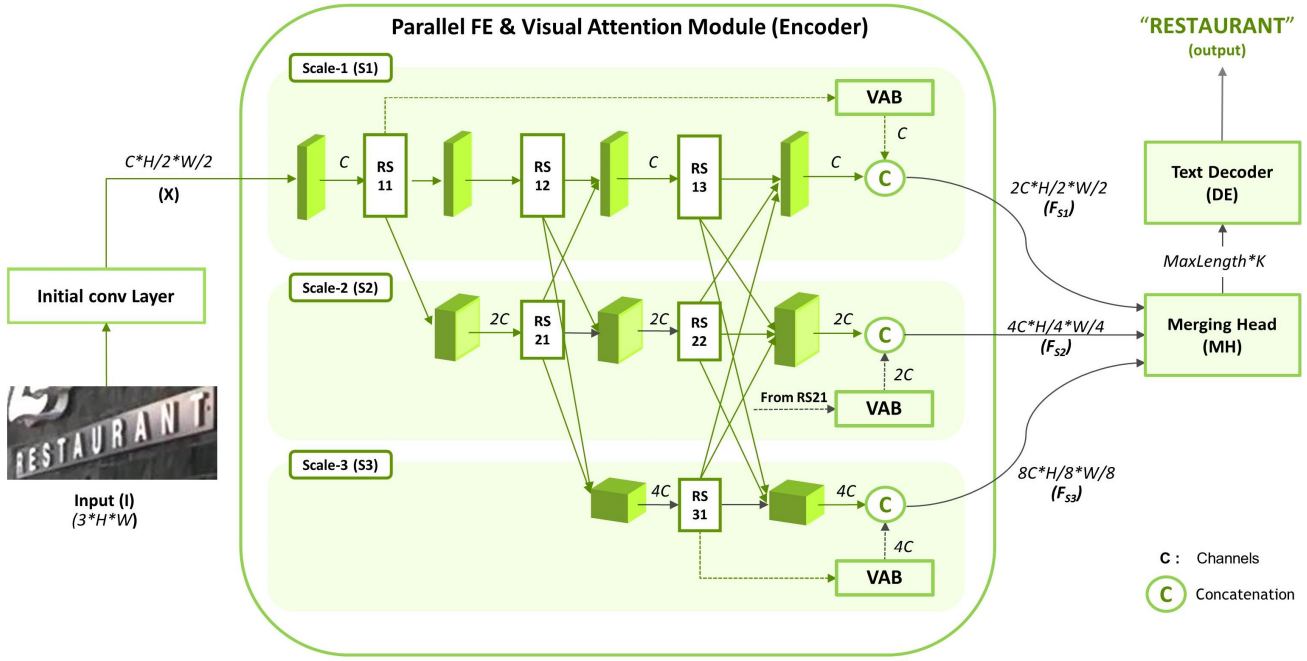


Fig. 2. The proposed text recognition network. Initially extracted features (X) from the input image (I) first pass through the multi-scale feature-extraction (FE) and visual attention (VA) based encoder (EN). The encoder performs both intra-scale processing and inter-scale fusion between three scale-modules (S1, S2, S3). During the intra-scale processing, the channels are processed with the residual connections-based residual structures (RS) as well as undergo the VA mechanism via the visual attention block (VAB), followed by their concatenation to produce the respective scale-module output. Consequently, the encoder outputs three sets of feature-maps that go through the merging head (MH) for channel and resolution adjustment. Finally, the text decoder (DE) outputs the recognized text character-wise.

these multi-scale modules also fuse their channels at regular intervals to develop the accumulated knowledge together. The encoder outputs three multi-scale channels (F_{S1}, F_{S2}, F_{S3}) that are merged together via the merging head (MH). The text decoder (DE) finally outputs the recognized text. The proposed network architecture consists of three major components: Text Encoder (EN), Merging Head (MH), and Text Decoder (DE) as detailed next.

A. Encoder (EN)

The purpose of the encoder is to simultaneously perform the feature extraction (FE) and visual attention/alignment (VA) on the multi-scale feature-maps. The input channels ($X \in \mathbb{R}^{C*\frac{H}{2}*\frac{W}{2}}$) pass through three multi-scale modules(S1, S2, S3) to finally yield three respective output feature-maps with different dimensions. The encoder processes the input feature-maps as follows:

$$(F_{S1}, F_{S2}, F_{S3}) = Encoder(X), \tag{1}$$

where $F_{S1} \in \mathbb{R}^{2C*\frac{H}{2}*\frac{W}{2}}$, $F_{S2} \in \mathbb{R}^{4C*\frac{H}{4}*\frac{W}{4}}$, $F_{S3} \in \mathbb{R}^{8C*\frac{H}{8}*\frac{W}{8}}$ and C indicates the total number of input channels. As we move from S1 to S3, the number of channels becomes twice as many as their subsequent upper scale. Similarly, the feature-map resolution (scale) decreases to half with each scale-module as we move from S1 to S3. It may be noted that each scale module keeps the channel resolution the same throughout that module [38], [40].

1) Intra-Scale Processing: Within every scale module (S1, S2, S3), the input channels pass through one or more residual structures (RS) and the visual attention process.

Residual Structure (RS). Each RS block comprises of five residual units (RU). The RU unit is a 3-layered residual building block as given in [10] that contains three convolution layers ($1 \times 1, 3 \times 3, 1 \times 1$) and a residual connection. After every convolution operation in the paper, we deploy the Batch-Normalization (BN) [12] and the ReLU activation [27] unless stated otherwise. The RS blocks are denoted as *RS(xy)*, where *x* denotes the scale-module number (1,2 or 3) and *y* indicates their location or index within that module (starting from left to right). Thus, *RS12* denotes the second RS block in the S1 scale-module.

Visual Attention Block (VAB). The scale-modules (S1, S2, S3) also visually attend (align) their feature-maps independently. This helps the network in making a better understanding regarding the foreground and background image pixels at different feature-scales. The first RS block output channels ($\in \mathbb{R}^{C'*H'*W'}$) in any scale-module undergo the attention mechanism via the *VAB* block. The attended feature-maps are then concatenated back at the end of the respective scale-module. The VAB process is shown in Fig. 3, where the input feature-maps first go through the five consecutive convolution layers. Next, a single feature-map is obtained via a simple 1×1 convolution operation. The sigmoid function is then applied on the resultant channel to obtain the segmentation map ($SM \in \mathbb{R}^{1*H'*W'}$). The SM undergoes element-wise

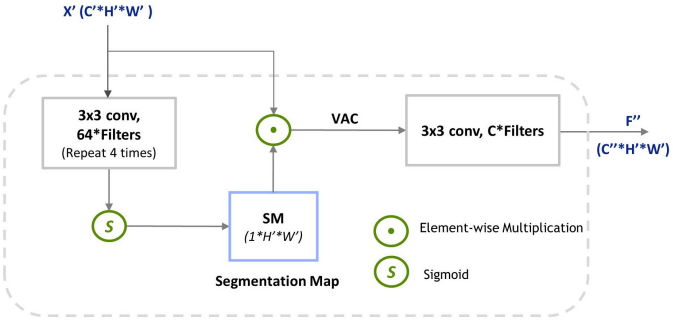


Fig. 3. Visual Attention Block (VAB). The input channels go through the convolution operation four repeated times. Subsequent single-filter 1×1 convolution and sigmoid function give the segmentation map (SM). The SM undergoes the element-wise multiplication with the original input channels to yield the visually attended channels (VAC) that are channel-adjusted to become the VAB output.

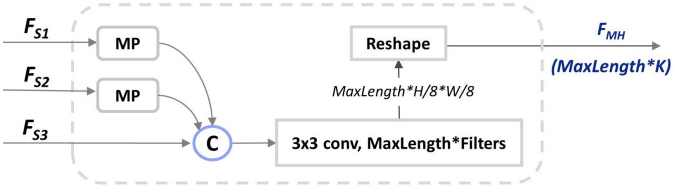


Fig. 4. Merging Head (MH). The higher scale-modules (S1,S2) output channels (F_{S1}, F_{S2}) are max-pooled before concatenation with the S3 scale-module output. Convolution and reshaping operations finally output the K-dimensional vectors with $MaxLength$ such vectors in total.

multiplication with the original input feature-maps to yield the visually-attended channels (VAC). The VAC feature-maps serve as the VAC module final output after being channel-adjusted via the 3×3 convolution operation. The VAB input feature-maps $X' \in \mathbb{R}^{C'*H'*W'}$ get visual attention as follows:

$$F'' = VAB(X'), \tag{2}$$

where $F'' \in \mathbb{R}^{C''*H'*W'}$ and we set $C'' = C'$. This attention process is different from the conventional and complex convolutional and deconvolutional layers based mechanism [43], and proves to be more effective as demonstrated in the experiments Sec. IV.

2) Repetitive Inter-scale Fusion: Inspired by the high-resolution networks [38], [40], the scale-modules (S1, S2, S3) also fuse channels with each other on regular intervals. It enables the network to form the accumulated and coordinated knowledge from the multi-scale channels and learn the valuable information better. To fuse the higher-scale source channels into the lower-scale target feature-maps, they undergo the $(n+1)$ times 3×3 convolution operation (with stride: 2, padding: 1). Here n $(= 0, 1)$ denotes the number of scale-modules in-between the source and target scale-modules. Thus, fusion from S1 channels into S3 requires two such convolution operations on S1 scale feature-maps to down-scale them to the S3 scale. Similarly, the lower-to-higher scale fusion

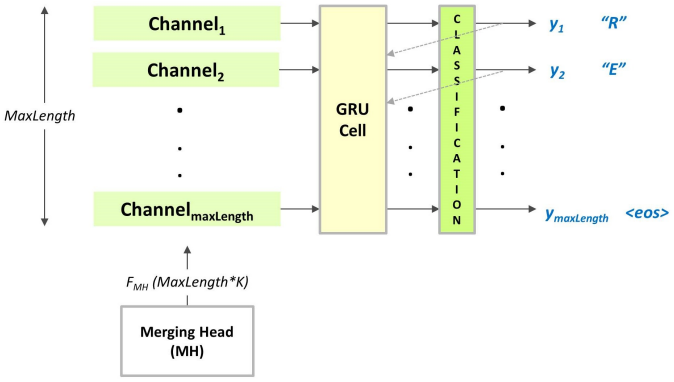


Fig. 5. Text Decoder (DE). The decoder predicts the recognized text character-by-character via the GRU and the classification layers. Here *eos* means the end-of-sequence character.

requires the bilinear upsampling of the lower-scale source feature-maps. No re-scaling transformation is done when the source and target scale-modules are the same. Once all source channels have been adjusted for channel quantity and target scale, they undergo the summation-based fusion with the target channels to obtain the fused feature-maps.

B. Merging Head (MH)

The encoder outputs three separate sets of feature-maps (F_{S1}, F_{S2}, F_{S3}) from the respective scale-modules (S1,S2,S3). The merging head (MH) combines them to output the feature-maps to be used for the text decoding. The MH block, as shown in Fig. 4, down-samples the S1 and S2 output channels using the max-pooling (MP) operation, so as to rescale them to the S3 output channels (F_{S3}) resolution. Next, they are concatenated together followed by the channel-adjustment via the convolutional layer. The resultant channels ($\in \mathbb{R}^{MaxLength*\frac{H}{8}*\frac{W}{8}}$) are reshaped into the K-dimensional vectors to give the MH final output ($F_{MH} \in \mathbb{R}^{MaxLength*K}$). Thus, the input channels are merged as follows:

$$F_{MH} = MH(F_{S1}, F_{S2}, F_{S3}), \tag{3}$$

Here, the MaxLength refers to the maximum length of text characters to be recognized. The output vectors are then routed to the text decoder for further processing.

C. Text Decoder (DE)

The responsibility of our text decoder is to perform recognition only. That makes it more focused on one task rather than the dual task of text alignment and recognition. We adopted the text decoder from the DAN network [43]. As shown in Fig. 5, the MH output channels ($F_{MH} \in \mathbb{R}^{MaxLength*K}$) go through the GRU [8] cell one-by-one at time ($t' = 1, 2, 3, ...MaxLength$) as K-dimensional vectors. The classification layer outputs the recognized text character at time t' with the output $p(y_{t'})$ as follows:

$$p(y_{t'}) = \text{softmax}(w * hidden_{t'} + b), \tag{4}$$

TABLE I

QUANTITATIVE EVALUATION ON THE STANDARD BENCHMARKS. THE RESULTS DEMONSTRATE THAT THE PROPOSED SCHEME IS THE MOST EFFECTIVE IN MOST CASES AS COMPARED TO THE SOTA METHODS ON THE RECOGNITION ACCURACY. THE BOLD AND UNDERLINED NUMBERS INDICATE THE BEST AND THE SECOND-BEST METHODS RESPECTIVELY.

	Rect.	Regular Datasets				Irregular Datasets		
Method		IIIT-5K	SVT	IC03	IC13	IC15	SVT-P	CUTE80
Jaderberg *et. al* [15]		-	80.7	93.3	90.8	-	-	-
Jaderberg *et. al* [13]		-	71.7	89.6	81.8	-	-	
Shi *et. al* [34]		81.2	82.7	91.9	89.6	-	-	-
Lyu *et. al* [24]		94.0	90.1	94.3	92.7	76.3	82.3	86.8
Xie *et. al* [44]		82.3	82.6	92.1	89.7	68.9	70.1	82.6
Liao *et. al* [20]		91.9	86.4	-	91.5	-	-	-
Cheng *et. al* [6]		87.4	85.9	94.2	93.3	70.6	-	-
Cheng *et. al* [7]		87.0	82.8	91.5	-	68.2	73.0	76.8
Bai *et. al* [3]		88.3	87.5	<u>94.6</u>	94.4	73.9	-	-
Yang *et. al* [46]		-	-	-	-	-	75.8	69.3
Shi *et. al* [36]	✓	93.4	89.5	94.5	91.8	76.1	78.5	79.5
Zhan *et. al* [49]	✓			-	91.3	76.9	79.6	83.3
Yang *et. al* [45]		93.3	90.2	91.2	93.9	78.7	80.8	<u>87.5</u>
Li *et. al* [18]		91.5	84.5	-	91.0	69.2	76.4	83.3
Liao *et. al* [19]		93.9	90.6	-	95.3	77.3	82.2	**87.8**
Wang *et. al* [43]		94.3	89.2	**95.0**	93.9	74.5	80.0	84.4
Yu *et. al* [47]		<u>94.8</u>	**91.5**	-	<u>95.5</u>	<u>82.7</u>	<u>85.1</u>	**87.8**
Ours		**95.9**	<u>90.8</u>	<u>94.6</u>	**96.3**	**83.9**	**86.0**	86.9

where $hidden_{t'}$ denotes the GRU cell hidden state, given as follows:

$$hidden_{t'} = GRU((embd_{t'-1}, Channel_{t'}), hidden_{t'-1}), \tag{5}$$

where $embd_{t'-1}$ is the embedding belonging to the previous classification $y_{t'-1}$. The network loss function is defined as follows:

$$L = -\sum_{t'=1}^{T'} \log P(y'_{t'}|Input, \theta) \tag{6}$$

where P indicates the prediction probability, $y'_{t'}$ is the actual or ground-truth text character at time t' and θ denotes the learnable parameters of the network.

IV. QUANTITATIVE AND QUALITATIVE EVALUATION

This section deals with the experimental analysis and comparison of the proposed network. First, we discuss the quantitative evaluation on seven standard benchmark datasets followed by the ablation study. We conclude with the visual analysis.

A. Experiments on Standard Benchmarks

Datasets. To evaluate the efficacy of the proposed network, we test on seven different scene-text datasets. They are either regular (IIIT-5k [25], IC03 [22], IC13 [17], SVT [41]) or irregular (IC15 [16], SVT-P [29], CUTE80 [30]) scene-text datasets.

IIIT-5k [25] is an internet-based scene-text dataset that contains 3,000 cropped text images for testing.

Street View Text (SVT) [41] comprises of 647 text-based test images collected via Google Street View. For diversity and variation, drastic corruption has been incorporated in the form of noise, blurriness, and low resolution.

ICDAR 2003 (IC03) [22] has 251 scene-text images with 867 test bounding boxes. As per the standard protocol [41], 860 cropped images have been retained after removing words with non-alphanumeric or less than 3 characters.

ICDAR 2013 (IC13) [17] is a regular scene-text dataset that contains total 1,015 cropped images, and most of them come from the IC03 dataset. Using the standard practice as given in [41], images with non-alphanumeric or less than three characters have been filtered out.

ICDAR 2015 (IC15) [16] contains irregular scene-text images taken via the Google Glasses with slight focusing and positioning. Only 1,811 test images have been utilized after removing some with extreme distortions as part of the standard pre-processing practice [6].

SVT-P [29] is an irregular scene-text dataset with 639 cropped images taken from Google street view. Mostly, they are single-angle based and highly perspective-distorted images.

CUTE80 [30] mainly deals with curved scene-text and consists of 80 images. We cropped 288 test samples from these high-resolution images using their bounding-box annotations.

Implementation Details. The input image gets resized with fixed height of 32 pixels and width up to 128 based on the aspect ratio. The proposed network is trained using two synthetic datasets until convergence: Synth90k [14] and SynthText [9]. A batch size of 64 has been used with 32 images each from Synth90k and SynthText. The value of total channels (C) in the encoder has been set to 32, so the scale-modules (S1, S2, S3) contain (32, 64, 128) channels respectively after every intra-scale processing step. MaxLength is set to 25, and the total number of character classes is 94 including the upper- and lower-case alphabets, 0-9 digits, and 32 ASCII punctuation symbols. The total decoder hidden units are set to 256. The

TABLE II
ABLATION STUDIES ON THE PROPOSED NETWORK. SEVERAL EXPERIMENTS ON DIFFERENT COMPONENTS OF THE PROPOSED NETWORK INDICATE THEIR VITALITY.

VAB Block Effect						
	IIIT5k	SVT	IC13	IC15	SVT-P	CUTE80
w/o VAB	86.6	81.9	88.8	77.4	80.7	75.2
w VAB (ours)	**95.9**	**90.8**	**96.3**	**83.9**	**86.0**	**86.9**
Number of Residual Units (RUs) per RS Block						
	IIIT5k	SVT	IC13	IC15	SVT-P	CUTE80
1	61.5	58.7	61.2	55.3	59.5	62.1
2	72.6	65.5	68.0	61.0	65.9	67.0
3	83.0	76.1	77.9	71.7	74.8	77.5
4	90.1	84.9	83.5	79.6	82.3	84.2
5 (ours)	**95.9**	**90.8**	**96.3**	83.9	**86.0**	**86.9**
6	94.3	88.8	95.6	**84.0**	85.5	86.2
S2 and S3 scale-modules Effect						
Scale-Modules	IIIT5k	SVT	IC13	IC15	SVT-P	CUTE80
S1 only	87.4	81.5	90.1	79.9	81.3	82.1
S1,S2 only	92.9	87.3	93.6	82.2	83.7	84.5
S1,S2,S3 (ours)	**95.9**	**90.8**	**96.3**	**83.9**	86.0	**86.9**
S1,S2,S3,S4	95.1	90.6	85.4	82.9	**86.5**	86.0
MaxLength Effect						
MaxLength	IIIT5k	SVT	IC13	IC15	SVT-P	CUTE80
25 (ours)	**95.9**	**90.8**	**96.3**	**83.9**	**86.0**	**86.9**
50	95.5	90.7	96.2	83.7	85.9	86.9
75	95.6	90.7	96.1	83.8	86.0	86.8
100	95.8	90.6	96.2	83.6	85.9	86.9

ADADELTA-based optimization [48] has been employed with the initial learning rate value of 1.0 and decreased to 0.1 from the fourth epoch.

Experimental Evaluation. Here, we compare our method quantitatively with the recent best networks. The comparison is done without using the lexicon information as it is generally the case in practice. As per the standard convention, the evaluation is done using the case-insensitivity for word accuracy computation. The results are shown in Table I, where our method outperforms other methods on 4 out of 7 datasets while performing reasonably competitive on the remaining three benchmarks. In comparison to the specifically designed rectification-based methods [23], [36], [49], our model gives better or competitive results without any rectification. For the regular scene-text dataset (IIIT-5K and IC13), we obtain an increase of (0.8% and 1.1%) respectively. While for the irregular scene-text datasets (IC15 and SVT-P), the proposed network improves the accuracy by (1.4% and 1.0%) respectively. The accuracy boost is mainly due to the inclusion of multi-scale visual attention and inter-scale fusion within the encoder. It is empirically shown during the ablation study as given in following paragraphs.

Ablation Study. We perform five different ablation experiments to analyze different components of our network.

1) *Effect of VAB Block*: The VAB block provides the most important visual attention mechanism that improves the network performance. As shown in Table II, the network underperforms on both regular and irregular scene-text datasets without using the VAB block. Thus, it's imperative to include the VAB block.

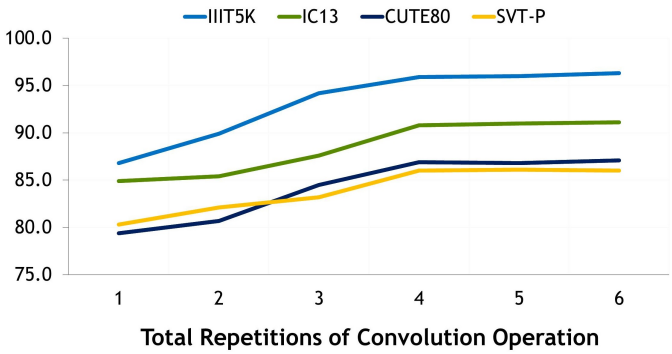


Fig. 6. VAB convolution quantity analysis graph. The graph indicates that repeating the convolution operation four times in the VAB block before the segmentation-map generation yields the optimal accuracy as tested on four different datasets.

2) *Number of Residual Units*: The number of residual units (RU) in the RS block plays an important role in better feature extraction. We experimented with different RU units quantity per RS block as shown in Table II. As per the results, we found five RU units per RS block to be the most effective choice with the highest accuracy.

3) *Effect of S2 and S3 scale-modules inclusion*: As given in Table II, using the S2 and S3 scale-modules in addition to S1 increases the network effectiveness. However, adding another scale-module S4 does not enhance the accuracy significantly. Thus, the (S1, S2, S3) combination has been employed.

4) *MaxLength Value Selection*: The MaxLength value has to be selected so that it covers the maximum length an output word can possibly have in a dataset. Beyond that, increasing it should not have any noticeable effect on the network efficacy. As given in Table II, increasing the MaxLength value from the default value of 25 does not alter the performance by much.

5) *Total Convolution Operations in VAB Block*: We investigate the effect of a total number of convolution operations before the segmentation map creation. To analyze the effect, we perform convolution operations quantity experiments on four datasets (IIIT5k, IC13, CUTE80, SVT-P). The results are shown in Fig. 6, where repeating four convolution operations before the segmentation map generation in the VAB block proves to be the best choice.

Robustness Analysis. Here, we check for robustness of the proposed scheme against different modifications on the input images. We compare our scheme with two recent SOTA methods (DAN [43] and CA-FCN [20]) on two datasets (IIIT-5K [25] and IC13 [17]). Following the practice as given in [43], variations introduced into these datasets are as follows:

IIIT-padded: 100% padding of the input images in IIIT-5k in both horizontal and vertical direction via border pixels replication. **IIIT-r-padded:** Stretching the image vertices using a random scale value up to 20% for both height and width respectively. Next, repetitive border pixels have been used for filling it. Finally, we crop the axis-aligned rectangles. **IC13-expansion:** The input images in IC13 are expanded into image frames with relatively extra 10% height and width followed

TABLE III

ROBUSTNESS ANALYSIS. THIS STUDY DEMONSTRATES THE PROPOSED MODEL HAS BETTER ROBUSTNESS TOWARDS DIFFERENT CHANGES IN THE INPUT IMAGES. (ACC: ACCURACY, DIFF.: ACCURACY DIFFERENCE OF PERFORMANCE FROM THE ORIGINAL DATASET, CHANGE (%): PERCENTAGE CHANGE (DECREASE) IN THE ACCURACY).

	IIIT	IIIT-padded			IIIT-r-padded			IC13	IC13-expanded			IC13-r-expanded		
Method	acc	acc	diff.	change (%)	acc	diff.	change (%)	acc	acc	diff.	change (%)	acc	diff.	change (%)
CA-FCN [20]	92.0	89.3	-2.7	2.9	87.6	**-4.4**	4.8	91.4	87.2	-3.7	4.1	83.8	-6.9	7.6
DAN-1D [43]	93.3	91.5	-1.8	1.9	88.2	-5.1	5.4	94.2	91.2	**-3.0**	**3.2**	86.9	-7.3	7.7
DAN-2D [43]	94.3	92.1	-2.2	2.3	89.1	-5.2	5.5	93.9	90.4	-3.5	3.7	86.9	-7.0	7.5
Ours	**95.9**	**94.0**	**-1.9**	**2.0**	**91.4**	-4.5	**4.7**	**96.3**	**92.5**	-3.8	3.9	**89.5**	**-6.8**	**7.1**

by cropping. **IIIT-r-expansion:** Expansion of the IC13 images using a random scale up to 20% height and width, followed by cropping the axis-aligned rectangular images.

As shown in Table III, it can be observed that the proposed method appears as the most stable and resilient to these input distortions and variations in majority cases, hence, demonstrating the robustness of our scheme.

B. Qualitative Analysis

Here, we present some good and bad qualitative results. We evaluate the proposed scheme with and without the visual attention block (VAB). The results are shown in Fig. 7, where the first two rows indicate the good results followed by the failure cases in the last row. Following the practice in [47], under each image, the first line shows the text recognition made by the proposed scheme without using the VAB block followed by our network text prediction with the VAB module in the second line. Characters colored as red indicate wrong predictions. As shown in the good results, the proposed scheme without the VAB block lacks the visual attention and struggles to differentiate between highly similar characters (e.g. 'e' and 'c' or 'o' and 'a') when they lack clear visual exposure, skewed perspective, or partial occlusion. The VAB block coupled with the multi-scale fusion helps in overcoming these issues and produces accurate results as shown.

The bad results, as shown in the last row of Fig. 7, mainly occur when the visual attention does not align the characters perfectly and results in failure as compared to the ground-truth (GT) recognition text.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new multi-scale and scale-wise visually attended text recognition network to address key scene-text challenges. The multi-scale feature extraction and visual attention have been performed in parallel to utilize different feature scales explicitly in a more effective way. The network also undergoes multi-scale fusion with each other to develop the coordinated information. Experimental evaluation on standard benchmarks indicates better accuracy in most cases as compared to the SOTA methods.

One of the key limitations of the proposed network is that it is using simpler inter-scale fusion. In the future, we aim to investigate more sophisticated fusion techniques. Moreover, our current work focuses only on the offline recognizer design, we will investigate the efficiency and computational cost aspects for real-time and real-world applications in the future.

Fig. 7. Ground truth (GT) scene-text based qualitative comparison. The first two rows demonstrate the good prediction results followed by the bad recognition cases in the last row. Under each image, the first line indicates our network text prediction without using the VAB block, whereas the second line shows our model with the VAB. The red-colored characters indicate the wrong predictions.

VI. ACKNOWLEDGEMENTS

This work was done while Usman was interning at SIE Global R&D. T. Kim was supported in part by the National Science Foundation (NSF) under grants CNS1955561 and AST2037864. G. Wang is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under grant RGPIN-2021-04244.

REFERENCES

[1] B. Su and S. Lu, "Accurate recognition of words in scenes without character segmentation using recurrent neural network," *Pattern Recognition*, vol. 63, pp. 397–405, 2017.

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[3] 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.

[4] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven, "Photoocr: Reading text in uncontrolled conditions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 785–792.

[5] R. G. Casey and E. Lecolinet, "A survey of methods and strategies in character segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 690–706, 1996.

[6] 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.

[7] 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.

[8] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[9] 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.

[10] 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, pp. 770–778.

[11] P. He, W. Huang, Y. Qiao, C. Loy, and X. Tang, "Reading scene text in deep convolutional sequences," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[13] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep structured output learning for unconstrained text recognition," *arXiv preprint arXiv:1412.5903*, 2014.

[14] Jaderberg, Max and Simonyan, Karen and Vedaldi, Andrea and Zisserman, Andrew, "Synthetic data and artificial neural networks for natural scene text recognition," *arXiv preprint arXiv:1406.2227*, 2014.

[15] Jaderberg, Max and Simonyan, Karen and Vedaldi, Andrea and Zisserman, Andrew, "Reading text in the wild with convolutional neural networks," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 1–20, 2016.

[16] 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 *13th IEEE International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 1156–1160.

[17] 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 *12th IEEE International Conference on Document Analysis and Recognition*, 2013, pp. 1484–1493.

[18] 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, 2019, pp. 8610–8617.

[19] M. Liao, P. Lyu, M. He, C. Yao, W. Wu, and X. Bai, "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[20] 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, 2019, pp. 8714–8721.

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[22] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto *et al.*, "Icdar 2003 robust reading competitions: entries, results, and future directions," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 7, no. 2-3, pp. 105–122, 2005.

[23] C. Luo, L. Jin, and Z. Sun, "Moran: A multi-object rectified attention network for scene text recognition," *Pattern Recognition*, vol. 90, pp. 109–118, 2019.

[24] P. Lyu, Z. Yang, X. Leng, X. Wu, R. Li, and X. Shen, "2d attentional irregular scene text recognizer," *arXiv preprint arXiv:1906.05708*, 2019.

[25] A. Mishra, K. Alahari, and C. Jawahar, "Scene text recognition using higher order language priors," in *BMVC-British Machine Vision Conference*, 2012.

[26] X. Mo, U. Sajid, and G. Wang, "Stereo frustums: a siamese pipeline for 3d object detection," *Journal of Intelligent & Robotic Systems*, vol. 101, no. 1, pp. 1–15, 2021.

[27] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[28] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3538–3545.

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

[30] 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.

[31] U. Sajid, W. Ma, and G. Wang, "Multi-resolution fusion and multi-scale input priors based crowd counting," *arXiv preprint arXiv:2010.01664*, 2020.

[32] U. Sajid, H. Sajid, H. Wang, and G. Wang, "Zoomcount: A zooming mechanism for crowd counting in static images," *IEEE Transactions on Circuits and Systems for Video Technology,*, vol. 30, no. 10, pp. 3499–3512, 2020.

[33] U. Sajid and G. Wang, "Plug-and-play rescaling based crowd counting in static images," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2287–2296.

[34] 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.

[35] 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.

[36] 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, 2018.

[37] B. Su and S. Lu, "Accurate scene text recognition based on recurrent neural network," in *Springer Asian Conference on Computer Vision*, 2014, pp. 35–48.

[38] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[40] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

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

[42] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in *Proceedings of the 21st IEEE International Conference on Pattern Recognition*, 2012, pp. 3304–3308.

[43] T. Wang, Y. Zhu, L. Jin, C. Luo, X. Chen, Y. Wu, Q. Wang, and M. Cai, "Decoupled attention network for text recognition." in *AAAI*, 2020, pp. 12 216–12 224.

[44] Z. Xie, Y. Huang, Y. Zhu, L. Jin, Y. Liu, and L. Xie, "Aggregation cross-entropy for sequence recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6538–6547.

[45] 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 International Conference on Computer Vision*, 2019, pp. 9147–9156.

[46] X. Yang, D. He, Z. Zhou, D. Kifer, and C. L. Giles, "Learning to read irregular text with attention mechanisms." in *IJCAI*, vol. 1, no. 2, 2017, p. 3.

[47] D. Yu, X. Li, C. Zhang, T. Liu, J. Han, J. Liu, and E. Ding, "Towards accurate scene text recognition with semantic reasoning networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 113–12 122.

[48] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[49] 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.