Stratified Multi-Task Learning for Robust Spotting of Scene Texts

Kinjal Dasgupta*
Department of Electrical Engineering
Heritage Institute of Technology, Kolkata
kinjal.dasgupta.9@gmail.com

Sudip Das*
CVPR Unit
Indian Statistical Institute, Kolkata
d.sudip47@gmail.com

Ujjwal Bhattacharya
CVPR Unit
Indian Statistical Institute, Kolkata
ujjwal@isical.ac.in

***Abstract*—Gaining control over the dynamics of multi-task learning should help to unlock the potential of the deep network to a great extent. In the existing multi-task learning (MTL) approaches of deep network, all the parameters of its feature encoding part are subjected to adjustments corresponding to each of the underlying sub-tasks. On the other hand, different functional areas of the human brain are responsible for distinct functions such as the Broca's area of the cerebrum is responsible for speech formation whereas its Wernicke's area is related to the language development etc. Inspired by this fact, in the present study, we used here a *Feature Representation Block* (FRB) of connection weights spanned over a few successive layers of a deep multi-task learning architecture and stratify the same into distinct subsets for their adjustments exclusively corresponding to different sub-tasks. Additionally, we have introduced a novel regularization component for controlled training of this *Feature Representation Block*. The purpose of the development of this learning framework is efficient end-to-end recognition of scene texts. Simulation results of the proposed strategy on various benchmark scene text datasets such as *ICDAR 2015*, *ICDAR 2017 MLT*, *COCO-Text* and *MSRA-TD500* have improved respective SOTA performance.**

***Index Terms*—Scene Text Spotting, Deep Learning, Multi-task Learning**

I. Introduction

Although text spotting in a natural scene image is a challenging task, its application potential is enormous. Some of its major applications include assistance to the blind or foreigners, surveillance, autonomous driving etc. The difficulties of this task are mainly due to the presence of texts of widely variable patterns against various complex backgrounds often having low contrast and possibly affected by non-uniform illumination, noise, occlusion etc. Moreover, such an image may contain important texts which are out of focus. On the other hand, spotting of such texts should be helpful for an automatic understanding of natural scenes. With the rapid advancements of deep learning methodologies, end-to-end text spotting from natural scene images has been an emerging research topic in the Computer Vision community. In such an attempt [1], [2], [3], detection of regions of scene image containing texts, its segmentation and finally recognition are performed successively. Sharing of information between respective processes of execution of these sub-tasks should be a tricky strategy for a successful text spotter.

*Indicates equal contributions

Figure 1. Text spotting result of the proposed framework on samples of *ICDAR 2015* (left) and *COCO-Text* (right) datasets. Red rectangles show detected text regions while texts printed in red color show recognized texts.

Usually, real-life problems like text spotting in natural scenes are composed of multiple tasks. Each of the sub-tasks may be performed in some independent manner such that execution of an individual sub-task does not consider any clue from the execution of the previous tasks possibly excluding the final outputs of the earlier tasks in the pipeline. It may also be accomplished in another way where all the tasks are performed in a collaborative manner shredding off their non-collaboration or mutual independence. Intuitively, the latter approach should be more efficient because the individual tasks are only sub-tasks of the given task. In fact, independent performance of individual tasks may require more effort while their outputs may suffer from incompatibility. For example, if the different pieces of some formal wear are collected independently, more time will be needed for the procurement and the pieces so collected may not be matching each other. Similarly, in the case of machine learning, multi-task learning is intuitively a more efficient approach than independent learning of multiple single tasks. This has the advantage of better generalization and requires a lesser volume of labelled training samples. So, researchers had studied multi-task learning since quite early days [4] of the study of machine learning. However, despite the remarkable success of the applications of single-task learning in different areas, efficient learning of multiple sub-tasks simultaneously remains a challenge to the Machine Learning community. A survey of existing approaches of multi-task learning can be found in [5].

In a multi-task scenario, machine learning algorithms deal with multiple loss functions, one corresponding to each of the sub-tasks while implementation of the same using a deep

network shares its lower few layers (where feature learning takes place) across different sub-tasks. This in effect helps in a better generalization of the resulting trained network. Thus, the learning experiences of individual sub-tasks are mutually exploited for generalized and fast learning of the underlying task. On the other hand, it causes the loss of sub-task specific feature information in the lower layers of the network affecting its overall recognition performance. Also, such an arrangement does not agree with the functional organization of human brain capable of performing multiple tasks in a very efficient way. It has distinct areas responsible for different tasks such as the occipital lobe is responsible for vision, the cerebellum regulates our motor movements like balance, posture, coordination etc., the temporal region takes care of our listening and so on. A similar strategy is adopted here by placing a *Feature Representation Block* (FRB) [6] consisting of multiple layers of connection weights at the top of common feature encoding network. A distinct subset of parameters of this FRB is selected dynamically for each of the underlying tasks such that the parameters belonging to the selected subset are subsequently adjusted based on the loss function of the corresponding task. Also, the novel regularizer proposed by us provides certain control over parameter learning of individual sub-tasks by means of neural inhibition.

Here, it is relevant to note that the sub-network responsible for text detection module of the entire recognition procedure had been used before in an earlier study [6] of the detection task alone. However, the text detection performance of the present multi-task learning framework has improved the earlier performance of the stand-alone detection network. Section IV-D of this article will provide further discussions on this issue.

Contributions of the present study are as follows.

1) A novel end-to-end trainable framework towards robust detection, segmentation and recognition of scene texts.
2) Introduction of a novel concept of parameter adjustments of FRB in a stratified manner by borrowing the idea from the functional organization of human brain. Also certain control on the parameter adjustments of different strata of the FRB has been achieved by introducing a novel regulalizer in the loss function of multi-task learning.
3) Comparative simulation studies on various benchmark sample databases including *ICDAR 2015*, *ICDAR 2017 MLT*, *COCO-Text* and *MSRA-TD500* establishes the improvement of the SOTA.

II. Related Works

Detection and recognition of texts in scene images have been studied for a long period. Here, we present a brief review of similar works. Existing detection and recognition approaches of scene texts have been recently analyzed in [7].

A. Text Detection

Both sliding window based [8], [9], [10] and connected component based [11], [12], [13] methods had been used in comparatively earlier studies of scene text detection. However, later studies are mostly based on various deep neural network architectures. In 2015, *ICDAR* organizers invited the Robust Reading Competition [14] based on a new dataset containing incidental scene text samples. In the following year, a similar dataset, called *COCO-Text* [15] was introduced. A fully connected neural network trained by Zhou *et al*. [16] can directly estimate the bounding boxes of incidental scene texts of arbitrary orientations and varying shapes. He *et al*. [17] proposed a direct regression strategy to draw quadrilateral boundaries of incidental texts of multiple orientations, widely variable size and often suffered by perspective distortions. In [18], a pyramid context network had been proposed for localization of text regions in natural scene images. TextSnake [19] introduced certain text instance representation strategy in a sequence of ordered, overlapping disks for efficient identification of horizontal, oriented or curved texts. Wang *et al*. [20] adopted region proposal network (RPN) for detection of scene texts of arbitrary shapes.

B. Text Recognition

The main objective of the text recognition system is the conversion of the variable-length cropped text images into machine-encoded text. Traditional methods [21], [22] of scene text recognition problem captures individual characters from the images and later refine the misclassified characters. In [23] authors, applied Histogram Oriented Gradient (HOG) to transform the word image toward the column vectors, later RNN is adapted to identify the corresponding word form the vectors. Jaderberg *et al*. [24] proposed a scene text recognition framework which does not require any human-annotated data. Zhan *et al*. [25] handles text recognition problem by using a sequence-to-sequence model with an attention mechanism together with a rectification module which helps to develop the accurate scene text iteratively effected from perspective distortions and curves. An end-to-end trainable neural network proposed in [26] which comprises of a rectification module based on Thin Plate Spine transformation along with Spatial Transformer Network and a sequence-to-sequence model extended by the bidirectional decoder for recognition purpose. Symmetry-constrained Rectification Network [27] has recently been used to recognize irregularly shaped text instances.

C. Text Spotting

Text spotting algorithms aim to detect texts from the scene images and then recognize those textual content from the cropped text patches via a recognizer model. Jaderberg *et al*. [28] proposed a region proposal technique for text detection purpose and a deep convolutional neural network for text recognition model. TextBoxes [29] handles the text detection problem by using an end-to-end trainable neural network later recognition achieved by connecting a CTC based recognizer module. Li *et al*. [1] proposed an end-to-end text spotter which consists of a text proposal network motivated by Region Proposal Network (RPN) and attention-based RNN decoder for recognition purpose. Liu *et al*. [2] computed low-level and high-level semantic feature maps with the help of their CNN to

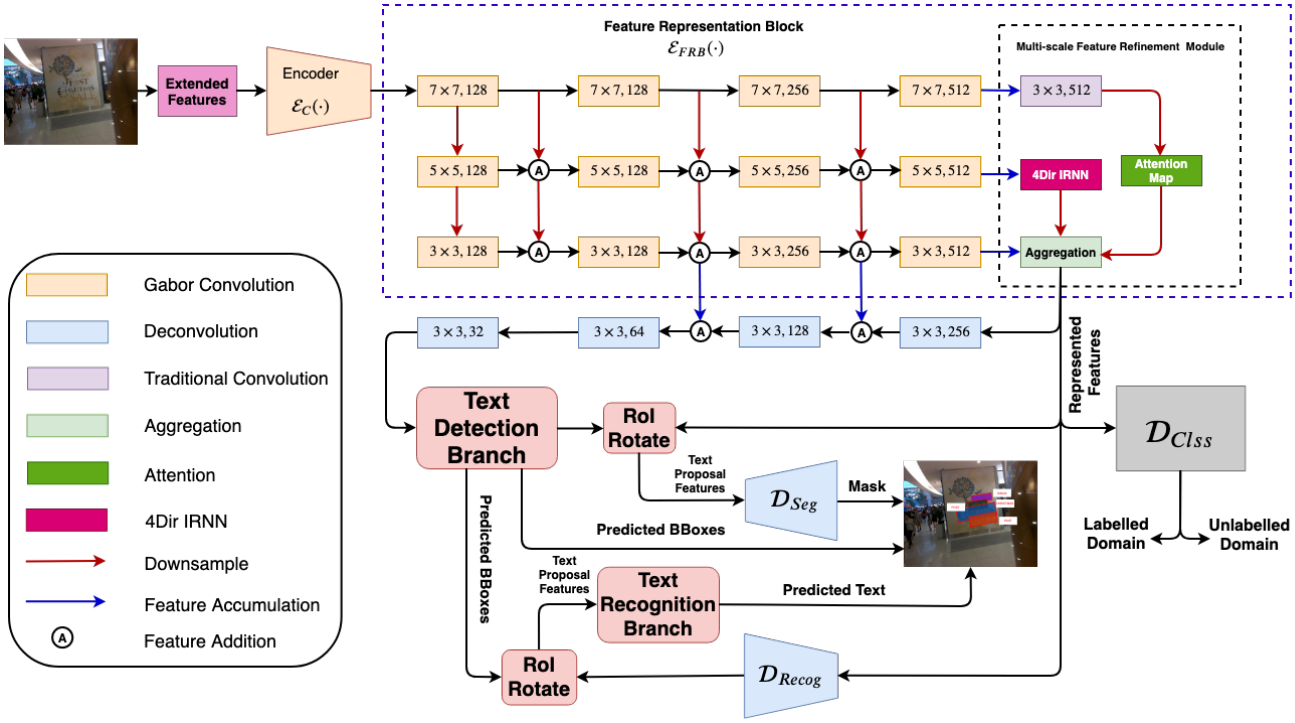

Figure 2. Schematic diagram of the proposed architecture: It contains (i) A ResNet-18 as an initial feature encoder $\mathcal{E}_c(\cdot)$, (ii) A *Feature Representation Block* (FRB) $\mathcal{E}_{FRB}(\cdot)$ consisting of several Gabor convolutional operators of different sizes arranged in three rows and a *Multi-scale Feature Refinement Module* (MFRM), (iii) A series of 4 deconvolution layers, (iv) A text detection branch, (v) A text segmentation branch $\mathcal{D}_{Seg}$, (vi) A domain adaptation module $\mathcal{D}_{Clss}$ and the (vii) Text recognition branch. The MFRM consists of a 4Dir IRNN, a channel-wise attention block and a CRFs-based aggregation block.

detect differently orientated incidental texts and later recognize text labels using the region features extracted by shared convolutions and transformed by RoIRotate. In [3] Mask R-CNN has been used for arbitrarily shaped text detection and a sequence-to-sequence model for text recognition.

III. Proposed Methodology

Details of feature encoding from the input image have been described in subsection III-A while the proposed multi-task learning architecture responsible for the next three successive main tasks (i) text detection, (ii) text recognition and (iii) text segmentation have been provided in the subsections III-B, III-C and III-D respectively. The training strategy of the proposed architecture is described later in subsection III-E.

A. Feature Extraction

This initial part of the network takes care of feature encoding from the input image. As shown in Figure 2, the initial block of feature encoding part causes extension of the input image to 4 orientation channels as in [30] and the next block is a ResNet-18 [31] (considered up to $Conv_{3\_2}$ layer) architecture modulated by Gabor orientation filters (GoFs) [30] which is further next followed by a *Feature Representation Block* (FRB) consisting of Gabor convolutional layers [30] arranged in multiple rows for enhancement of their traditional counterparts with respect to various image transformations such as scale variations, rotations etc. and another distinct sub-block for abstract feature extraction, termed as *Multi-scale Feature Refinement Module* (MFRM), consisting of a 4Dir IRNN [32] for extraction of contextual features and a channel-wise attention map which enhances important features while attenuates noisy unwanted background information. FRB output is finally emanated through an aggregation block consisting of conditional random fields (CRFs) [33].

B. Detection

The above output from the FRB is fed to a decoder block consisting of a series of deconvolution layers before feeding the detection branch (detail configuration avoided in Figure 2 due to space restriction) placed next to the last of four deconvolution layers. This part of the network, designed for instant scene text detection, consists of three $Conv_{1\times1}$ layers operated in parallel generating the output consisting of a single-channel score map and the RBOX geometry which in turn consists of 4 channels for each text box and a single-channel towards its orientation angle. This detection branch uses the following loss function.

$$\mathcal{L}_{Detect} = \mathcal{L}_S + \lambda_G \mathcal{L}_G, \tag{1}$$

where $\mathcal{L}_S$ and $\mathcal{L}_G$ are the two components of the loss function corresponding to the Score Map and the Geometry respectively while the hyper-parameter λ_G plays the role of balancing agent between the two losses.

Table I
NETWORK CONFIGURATION OF THE TEXT RECOGNITION MODULE. EACH BLOCK FROM THE FIRST ROW ONWARD IS A RESIDUAL BLOCK. $\mathcal{S}$ DENOTES STRIDE OF THE FIRST CONVOLUTIONAL LAYER OF EACH BLOCK.

	Layers	Feature Maps
Encoder	$\mathcal{C}onv_{7\times7}$, $\mathcal{S}_{1\times1}$	32
	$\mathcal{C}onv_{3\times3}$, $\mathcal{S}_{1\times1}$	32
	$\begin{bmatrix}\mathcal{C}onv_{1\times1}\\ \mathcal{C}onv_{3\times3}\end{bmatrix}\times3$, $\mathcal{S}_{1\times1}$	32
	$\begin{bmatrix}\mathcal{C}onv_{1\times1}\\ \mathcal{C}onv_{3\times3}\end{bmatrix}\times4$, $\mathcal{S}_{2\times2}$	64
	$\begin{bmatrix}\mathcal{C}onv_{1\times1}\\ \mathcal{C}onv_{3\times3}\end{bmatrix}\times6$, $\mathcal{S}_{2\times1}$	128
	$\begin{bmatrix}\mathcal{C}onv_{1\times1}\\ \mathcal{C}onv_{3\times3}\end{bmatrix}\times3$, $\mathcal{S}_{2\times1}$	256
	BiLSTM 1 BiLSTM 2	256 256
Decoder	Attentional LSTM	256 Attention Units, 256 Hidden Units
	Attentional LSTM	256 Attention Units, 256 Hidden Units

C. Recognition

The output from the FRB is passed to the RoIRotate [2] after passing through $\mathcal{D}_{Recog}$ consisting of a series of deconvolution layers. $\mathcal{D}_{Recog}$ reduces the number of input feature maps while increasing their resolution to facilitate robust recognition of small sized texts. RoIRotate also receives the RBOX geometry information from the text detection branch. It horizontally aligns the upsampled feature map obtained from $\mathcal{D}_{Recog}$ by rotating it using the orientation angle of the RBOX information and also resizes it. This output of RoIRotate is passed to the Text recognition branch to predict the sequence of characters in the detected text box.

Configuration of the text recognition branch is shown in Table I. Its encoder part consists of a 34-layer residual network and two Bidirectional LSTM (BiLSTM) layers. Each of its residual units performs two successive convolution operations of size 1×1 and 3×3 respectively. Downsampling of feature maps is done as in [26]. The decoder part consists of a 2-layer attentional LSTMs each of which comprises of 256 hidden units and another 256 attention units. The loss function used for the recognition branch is similar to [2].

D. Unsupervised Segmentation

The later stage of the proposed framework is a text segmentation network that is trained to estimate the segmentation mask of the entire text instances in an unsupervised manner. Since segmentation map of the text instances for all of the benchmark datasets are not available excluding *COCO-Text*, we take a step further to use Unsupervised Adversarial Domain Adaptation [34] to estimate the segmentation map with the help of an additional *COCO-Text* dataset.

As shown in the Figure 2, the proposed approach for the part of the segmentation task consists of common encoder (ResNet-18 $\mathcal{E}_C(\cdot)$ followed by a FRB $\mathcal{E}_{FRB}(\cdot)$), a decoder network $\mathcal{D}_{Seg}$ and a domain classifier $\mathcal{D}_{Clss}$. The encoder takes the image as input to extract the dense representation of the features. RoIRotate [2] crops the RoIs from the extracted feature maps of FRB and scales it to a fixed size feature map which is proposed by the detection branch and then it is passed through a decoder network to estimate the segmentation map precisely. The pair of the {Domain classifier, Encoder} and {Encoder, Decoder} networks are trained alternatively in an adversarial manner to reduce the domain shift between *COCO-Text* and other datasets. We employ an adversarial approach to handle the problem of domain transfer between datasets, encoder (ResNet-18 $\mathcal{E}_C(\cdot)$ followed by an FRB $\mathcal{E}_{FRB}(\cdot)$) is related to the Generator of a Generative Adversarial Network (GAN) [35] and $\mathcal{D}_{Clss}$ is comparable to the Discriminator. The common encoder here aims to generate dataset invariant features for input scene text instances from the two datasets while $\mathcal{D}_{Clss}$ aims to classify these features as to which domain they belonged to. The loss function is used in the domain adaptation approach and segmentation is as similar to [36].

E. Training Strategy

Usually, it is difficult to train a multi-task network [37]; training with respect to each individual task should be properly balanced so that the network parameters converge to certain robust shared features that are useful across all the tasks [37]. On the other hand, multi-task learning (MTL) based network is considered a superior model since here learning with respect to one task is assisted by the learning of the remaining tasks. Moreover, MTL across different related tasks not only takes advantage of cross-task information but also it has the advantage of enhanced generalization. Although the present task of scene text spotting consists of three successive sub-tasks of text detection ($\mathcal{T}^{Det}$), recognition ($\mathcal{T}^{Recog}$) and segmentation ($\mathcal{T}^{Seg}$), we consider here an additional sub-task of domain classification ($\mathcal{T}^{D_{Clss}}$) to reduce the domain shift between two different dataset distribution. The common encoder part of this MTL network is trained by all the above four sub-tasks. The loss function used for training of our MTL network includes a novel regularizer to achieve some control over parameter learning for each particular task utilizing neural inhibition.

Training of our MTL network has been performed in two phases as follows. In the first phase, we initialize the relevant network parameters separately for learning of each of the four individual sub-tasks $\mathcal{T}^{Det}, \mathcal{T}^{Recog}, \mathcal{T}^{Seg}$ and $\mathcal{T}^{D_{Clss}}$ independent of each others and store the respective sets of

values of network parameters of the *Feature Representation Block* (FRB) only. Thus, in the first phase, we train four sub-networks denoted by Θ^{Det}, Θ^{Recog}, Θ^{Seg} and $\Theta^{D_{Clss}}$ responsible for the above four sub-tasks respectively. A truncated normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{0.1})$ is used for initialization of the parameters of each sub-network. Here, Θ^{Det} excludes $\mathcal{D}_{Recog}$, $\mathcal{D}_{Seg}$ and $\mathcal{D}_{Clss}$ parts from the whole network presented in Figure 2. Similarly, the remaining three sub-networks do not include the respective irrelevant parts of the whole network of Figure 2. The above strategy has been taken to make the remaining tasks purposefully freeze during the single-task learning and we separately optimize the loss functions of respective sub-tasks.

In the second phase, we consider the whole network of Figure 2 for training of the underlying multi-task of scene text spotting. Initially, the parameters of the network are initialized using $\mathcal{N}(\mathbf{0}, \mathbf{0.1})$ as before. Here, parameter adjustment takes place with respect to the loss function $\mathcal{L}_{Multi}$ of equation 2.

$$\mathcal{L}_{Multi} = \mathcal{L}_{Detect} + \lambda_{Recog}\mathcal{L}_{Recog} + \lambda_{Seg}\mathcal{L}_{Seg} + \lambda_{D_{Clss}}\mathcal{L}_{D_{Clss}} + (\lambda/2)\mathcal{R}, \quad (2)$$

where λ_{Recog}, λ_{Seg}, $\lambda_{D_{Clss}}$ are respectively the hyper-parameters of the loss functions of the sub-tasks $\mathcal{T}^{Recog}$, $\mathcal{T}^{Seg}$, and $\mathcal{T}^{D_{Clss}}$, and λ is the hyper-parameter of the regularizer.

$$\mathcal{R} = \sum_i w_i[(\theta_i - \theta_i^{Det})^2 + (\theta_i - \theta_i^{Recog})^2 + (\theta_i - \theta_i^{Seg})^2 + (\theta_i - \theta_i^{D_{Clss}})^2], \quad (3)$$

where θ_i is a parameter of the multi-task network, $\theta_i^{Det} \in \Theta^{Det}, \theta_i^{Recog} \in \Theta^{Recog}, \theta_i^{Seg} \in \Theta^{Seg}, \theta_i^{D_{Clss}} \in \Theta^{D_{Clss}}$ and i ranges over the parameters of the FRB. w_i indicates the weighting importance of parameter θ_i for each $i \in$ FRB. However, similar to the functional organization of the human brain, we consider only a subset of the parameters of the FRB during each iteration for the adjustments. This subset is selected dynamically by computing the minimum distance of the current value of each parameter of the FRB from the respective values of the four sets of saved parameters corresponding to the four single-task learning. If this minimum value of a parameter of the FRB corresponds to the task $\mathcal{T}^t$, where t ranges over the four sub-tasks, then this parameter is adjusted using the derivative of the loss function of t-th sub-task.

IV. Experimentation Details

A. Datasets

ICDAR 2015 [14] dataset includes 1000 training images and 500 testing incidental scene images used for oriented scene text detection and spotting. Text in the scene can be in arbitrary orientations or suffer from motion blur and low resolution.

ICDAR 2017 MLT [41] is a large multi-lingual, multi-oriented, multi-script text dataset, includes 7200 training images, 1800 validation images and 9000 testing images.

MSRA-TD500 [42] is a dataset comprises 300 training images and 200 test images. Text areas are arbitrary orientations including line level annotations, contains text in both English plus Chinese.

COCO-Text [15] is a large dataset based on MS-COCO. The dataset contains a total 63,686 number of images, in which 43,686 are for the training purpose and remaining 20,000 are for validation and testing purposes.

B. Implementation Details

We use ResNet-18 before the FRB block as part of the encoder. Instead of ResNet-18 + FRB [6] blocks, we explore four more popular backbones as a single encoder, pre-trained on ImageNet [46]. Our ablation experiments in Section IV-D show that custom learning of weights in FRB significantly surpasses the network performance in terms of F-Score. Apart from that, we employ curriculum learning [47] concerning the complexity of images to enable the feature learning invariant in terms of degradation, different lighting conditions, and image complexities. Additionally, we used the *Mask and Predict* [6] strategy to enable the network to produce satisfactory results in cases of occlusions. The hyper-parameters λ, λ_G, λ_{Recog}, λ_{Seg} and $\lambda_{D_{Clss}}$ is set to 1 in our experiments. Momentum Optimizer is used with an initial learning rate of 0.01 and momentum of 0.9 during training. The learning rate is decreased by a factor of 10 after each subsequent 15k iterations. Data Augmentation is also employed to improve the robustness of the proposed framework. For data augmentation, we make use of random mirror and resize between 0.5 and 2 for all datasets, additional we add some random rotation between -10 to 10 degrees, Gaussian blur, brightness and contrast for each sample of the dataset. This comprehensive data augmentation scheme makes the network to resist overfitting and improve accuracy. The training of the network has been carried out on a computer with two NVIDIA P6 GPU.

C. Evaluation Results and Metrics

Results of our proposed framework shows that our model significantly outperforms on detection and recognition of the oriented and scale-invariant scene text on the benchmark *ICDAR 2015*, *ICDAR 2017 MLT*, *COCO-Text* and *MSRA-TD500* datasets. Among this *ICDAR 2015* and *ICDAR 2017 MLT* datasets contain some images which are suffered from blur, low-resolution and different lighting condition labelled as "DO NOT CARE". Experiments were done in [2] ignored image samples which are effected by blur during training. However, our model has been purposefully trained on the entire training datasets. We only show the detection result of *ICDAR 2017 MLT* and *MSRA-TD500* since the following two datasets do not have text spotting task. We use the *ICDAR* evaluation rules in terms of Recall, Precision, F-Score for the text spotting assessment purpose. A detected text region is identified as being corrected if the IOU with ground truth is larger than 0.5. F-Score is a particular standard of quality by consolidating recall and precision where recall and precision is the ratio of the number of precisely identified text regions

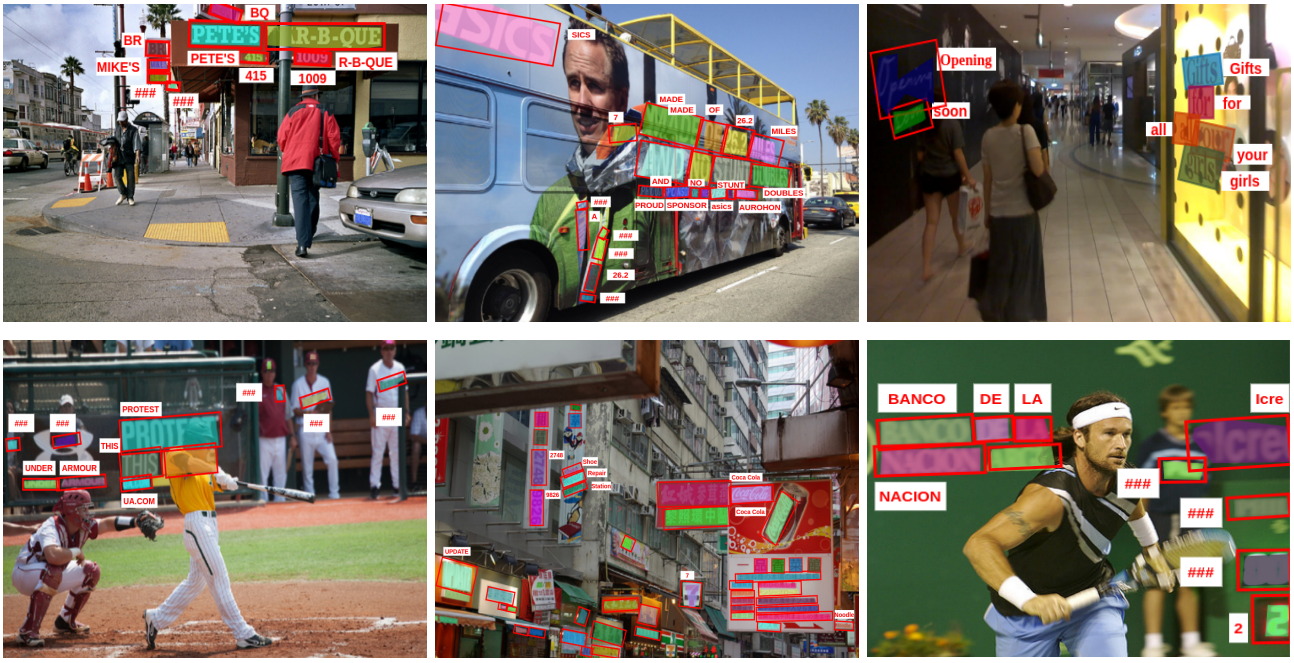

Figure 3. Text detection, segmentation and recognition performance (including a few failure cases) on scene images of *ICDAR 2015*, and *COCO-Text* datasets are shown. Red bounding box is used for detection result, highlighting is used for segmentation result while texts printed in red color against white background show recognized texts. Only English texts have been recognized. Wrongly recognized English texts are marked by the label #.

Table II
RESULTS OF DIFFERENT SOTA MODELS ON *ICDAR 2015* DATASET.

Methods	**Detection**			**Methods**	**End-to-End**		
	Recall	**Precision**	**F-Score**		**Strong**	**Weak**	**Generic**
EAST [16]	78.3	83.3	80.7	Stradvision-2 [14]	43.7	-	-
He *et al.* [38]	87.0	86.0	87.0	He *et al.* [38]	82.0	77.0	63.0
TextSnake [19]	84.9	80.4	82.6	Stradvision-1 [14]	33.2	-	-
MaskTextSpotter [39]	81.0	91.6	86.0	MaskTextSpotter [39]	79.3	73.0	62.4
FOTS [2]	87.9	91.8	89.8	FOTS [2]	83.5	79.1	65.3
Wang *et al.* [20]	86.0	89.2	87.6	NJU [14]	32.6	-	-
Qin *et al.* [3]	87.9	91.6	89.7	Qin *et al.* [3]	85.5	81.9	69.9
Dasgupta *et al.* [6]	89.2	91.3	90.2	TextNet [40]	78.6	74.9	60.4
Ours	**90.7**	**93.4**	**92.0**	**Ours**	**89.7**	**84.6**	**72.1**

Table III
RESULTS OF DIFFERENT SOTA MODELS ON *ICDAR 2017 MLT* DATASET.

Methods	**Detection**		
	Recall	**Precision**	**F-Score**
FOTS [2]	62.3	81.8	70.7
Xie *et al.* [18]	68.6	80.6	74.1
Textboxes [29]	62.1	45.5	52.5
Sensetime OCR [41]	69.4	56.9	62.5
SARI_ FDU_ RRPN_v1 [41]	55.5	71.1	62.3
NLPR-PAL [41]	57.9	76.6	66.0
Lyu *et al.* [43]	70.6	74.3	72.4
Dasgupta *et al.* [6]	73.9	88.6	80.5
Ours	**75.4**	**89.1**	**81.6**

Table IV
RESULTS OF DIFFERENT SOTA MODELS ON *MSRA-TD 500* DATASET.

Methods	**Detection**		
	Recall	**Precision**	**F-Score**
EAST [16]	67.4	87.3	76.1
Xie *et al.* [18]	68.6	80.6	74.1
Wang *et al.* [20]	**82.1**	85.2	83.6
Lyu *et al.* [43]	76.2	87.6	81.5
Dasgupta *et al.* [6]	81.6	**88.2**	**84.7**
Ours	81.6	**90.2**	**85.6**

Table V
DETECTION AND END-TO-END EVALUATION RESULTS ON *ICDAR 2015* DATASET OF DIFFERENT ENCODER AND LEARNING MODELS.

Learning Strategy	**VGG** [44]		**ResNet-50** [31]		**ResNet-101** [31]		**ResNeXt-101** [45]		**Ours**	
	F-Score$_{DET}$	**F-Score**$_{E2E}$	**F-Score**$_{DET}$	**F-Score**$_{E2E}$	**F-Score**$_{DER}$	**F-Score**$_{E2E}$	**F-Score**$_{DET}$	**F-Score**$_{E2E}$	**F-Score**$_{DET}$	**F-Score**$_{E2E}$
Naive Multi-task Learning	72.1	65.4	77.6	68.8	79.3	70.7	83.5	71.3	**83.8**	**73.0**
Stratified Multi-task Learning (- Curriculum Learning)	78.4	67.9	80.7	70.2	82.7	72.8	84.7	75.1	**89.6**	**85.8**
Stratified Multi-task Learning (+ Curriculum Learning)	80.2	70.5	83.7	74.5	85.1	76.6	88.4	82.3	**92.0**	**89.7**

Table VI
RESULTS OF DIFFERENT SOTA MODELS ON *COCO-Text* DATASET.

Methods	**Detection**			**End-to-End**
	Recall	**Precision**	**F-Score**	
EAST [16]	32.4	50.3	39.4	-
Lyu *et al*. [43]	32.4	61.9	42.5	-
Veit *et al*. [15]	23.3	**83.7**	36.4	40.0
Dasgupta *et al*. [6]	**50.6**	71.6	**59.2**	-
Ours	**53.4**	70.2	**60.6**	**48.7**

to the entire number of text regions within the dataset and the ratio of the number of precisely identified text regions to the total number of detected text regions. "S", "W", "G" present in the *ICDAR 2015* end-to-end evaluation protocol represents F-Score using "Strong", "Weak", "Generic" lexicon respectively. No lexicon is used in the end-to-end evaluation of *COCO-Text*. The comparative study on *ICDAR 2015* dataset is presented in Table II. Detection performance of the proposed model outperforms the state-of-the-art [6] model by almost 2% with respect to F-Score while end-to-end evaluation result improves the existing state-of-the-art [3] by about 4.9%.

Results of another comparative study on *ICDAR 2017 MLT* dataset is shown in Table III. The proposed model has improved the SOTA [6] on this dataset with respect to each of Precision, Recall and F-Score.

Comparative performance results on *MSRA-TD500* and *COCO-Text* datasets are shown in Table IV and Table VI respectively. Text detection performance of the proposed strategy has improved the existing SOTA [6] on both the datasets with respect to respective F-Score values.

Finally, evaluation of the performance of the proposed end-to-end model on *ICDAR 2015* and *COCO-Text* datasets shows highly significant improvements of the respective state-of-the-arts as shown in Table II and Table VI. Some of the failures of the proposed text spotting framework on complex scene images are shown in Figure 3.

D. Ablation Study

We have also performed several other simulations towards ablation study of the proposed framework. Towards the same, we used test samples of *ICDAR 2015* dataset and F-Score as the evaluation metric. Results of this experimentation have been shown in Table V. This Table includes results corresponding to the use of (i) VGG net [44], (ii) ResNet-50 [31], (iii) ResNet-101 [31], (iv) ResNeXt-101 [45] and (v) our scheme (ResNet-18 + FRB) as the feature encoder.

1) Baselines: In this study, we considered two baselines: detection only (DET) single-task and end-to-end ($E2E$) multi-task along with each of the following three backbone learning schemes. In case of (DET) baseline, we report detection performance while we report text spotting performance for ($E2E$) baseline. In each case, detection performance is better for multi-task learning ($E2E$) compared to (DET) learning.

2) Encoders: Our ablation study includes three encoder learning schemes: (i) Naive multi-task learning, (ii) Stratified multi-task learning without curriculum learning and (iii) Stratified learning coupled with curriculum learning. As we scan Table V horizontally from left to right, we can see the performance improve gradually and the best results have been obtained corresponding to our scheme (ResNet-18 + FRB) with respect to both single-task and multi-task learning schemes. Similarly, as we scan this table vertically from top to bottom, we observe that these results improve steadily and the highest performance could be achieved when the proposed stratified learning scheme gets coupled with the curriculum learning strategy.

V. CONCLUSIONS

In this article, we have presented details of our recent study of a novel multi-task learning framework designed for scene text spotting. Various sub-tasks performed by this network includes text detection, text segmentation and text recognition. Also, it has an additional module $\mathcal{D}_{Clss}$ which takes care of the much important domain adaptation task which helps the trained network to compensate from the shift in distributions of two different datasets of samples. Usually, such a multi-task framework developed using a deep architecture has a common feature encoding part, parameters of which are adjusted by training samples of all the sub-tasks. Thus, task specific features cannot be preserved into its feature encoder which in turn may affect the performance of the framework. The *Feature Representation Block* placed at the top of the feature encoding component of the network should take care of this issue. Its parameters are dynamically divided into a number of strata equal to the number of individual tasks such that the parameters of each stratum are adjusted by using the loss function of the corresponding task. Also, we have introduced a novel regulalizer component into the underlying multi-task loss function to control learning of each individual division of the *Feature Representation Block*. Introduction of these task specific parts into the *Feature Representation Block* has stem from the functional organization of human brain distinct parts of which are responsible for different functions performed by human beings. We have conducted extensive simulations of

the proposed framework on four benchmark datasets including *ICDAR 2015*, *ICDAR 2017 MLT*, *MSRA-TD500* and *COCO-Text* to perform detailed comparisons of results with the state-of-the-art models.

Present implementation of FRB stratification needs learning of each of the individual sub-tasks and storage of the corresponding learned parameter values of the FRB. However, this strategy may be replaced by a more sophisticated and efficient method. Future studies on this topic should concentrate in the above direction. Additionally, we may search for a better regularizer of the multi-task loss function as a part of the future studies.

REFERENCES

[1] H. Li, P. Wang, and C. Shen, "Towards end-to-end text spotting with convolutional recurrent neural networks," in *ICCV*, 2017, pp. 5238–5246.

[2] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "FOTS: Fast oriented text spotting with a unified network," in *CVPR*, 2018, pp. 5676–5685.

[3] S. Qin, A. Bissacco, M. Raptis, Y. Fujii, and Y. Xiao, "Towards unconstrained end-to-end text spotting," in *ICCV*, 2019, pp. 4704–4714.

[4] R. Caruana, "Multitask learning: A knowledge-based source of inductive bias," in *ICML*, 1993, pp. 41–48.

[5] Y. Zhang and Q. Yang, "A survey on multi-task learning," in *arXiv*, 2017.

[6] K. Dasgupta, S. Das, and U. Bhattacharya, "Scale-invariant multi-oriented text detection in wild scene image," in *ICIP*, 2020, pp. 2041–2045.

[7] Y. Zhu, C. Yao, and X. Bai, "Scene text detection and recognition: Recent advances and future trends," in *Frontiers of Computer Science*, vol. 10, no. 1, 2016, pp. 19–36.

[8] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in *ECCV*, 2014, pp. 512–528.

[9] S. Zhu and R. Zanibbi, "A text detection system for natural scenes with convolutional feature learning and cascaded classification," in *CVPR*, 2016, pp. 625–632.

[10] K. I. Kim, K. Jung, and J. H. Kim, "Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm," in *TPAMI*, vol. 25, no. 12, 2003, pp. 1631–1639.

[11] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *CVPR*, 2010, pp. 2963–2970.

[12] M. Busta, L. Neumann, and J. Matas, "FASText: Efficient unconstrained scene text detector," in *ICCV*, 2015, pp. 1206–1214.

[13] L. Neumann and J. Matas, "Scene text localization and recognition with oriented stroke detection," in *ICCV*, 2013, pp. 97–104.

[14] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou *et al.*, "ICDAR 2015 competition on robust reading," in *ICDAR*, 2015, pp. 1156–1160.

[15] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "COCO-Text: Dataset and benchmark for text detection and recognition in natural images," in *arXiv*, 2016.

[16] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "EAST: an efficient and accurate scene text detector," in *CVPR*, 2017, pp. 5551–5560.

[17] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Deep direct regression for multi-oriented scene text detection," in *ICCV*, 2017, pp. 745–753.

[18] E. Xie, Y. Zang, S. Shao, G. Yu, C. Yao, and G. Li, "Scene text detection with supervised pyramid context network," in *AAAI*, vol. 33, 2019, pp. 9038–9045.

[19] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, "TextSnake: A flexible representation for detecting text of arbitrary shapes," in *ECCV*, 2018, pp. 20–36.

[20] X. Wang, Y. Jiang, Z. Luo, C.-L. Liu, H. Choi, and S. Kim, "Arbitrary shape scene text detection with adaptive text region representation," in *CVPR*, 2019, pp. 6449–6458.

[21] C.-Y. Lee, A. Bhardwaj, W. Di, V. Jagadeesh, and R. Piramuthu, "Region-based discriminative feature pooling for scene text recognition," in *CVPR*, 2014, pp. 4050–4057.

[22] I. J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, and V. Shet, "Multi-digit number recognition from street view imagery using deep convolutional neural networks," in *arXiv*, 2013.

[23] B. Su and S. Lu, "Accurate scene text recognition based on recurrent neural network," in *ACCV*, 2014, pp. 35–48.

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

[25] F. Zhan and S. Lu, "Esir: End-to-end scene text recognition via iterative image rectification," in *CVPR*, 2019, pp. 2059–2068.

[26] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "Aster: An attentional scene text recognizer with flexible rectification," in *TPAMI*, vol. 41, no. 9, 2018, pp. 2035–2048.

[27] 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 *ICCV*, 2019, pp. 9147–9156.

[28] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," in *IJCV*, vol. 116, no. 1, 2016, pp. 1–20.

[29] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network," in *AAAI*, 2017, pp. 4161–4167.

[30] S. Luan, C. Chen, B. Zhang, J. Han, and J. Liu, "Gabor convolutional networks," in *TIP*, vol. 27, no. 9, 2018, pp. 4357–4366.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[32] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *CVPR*, 2016, pp. 2874–2883.

[33] L. Liu, Z. Qiu, G. Li, S. Liu, W. Ouyang, and L. Lin, "Crowd counting with deep structured scale integration network," in *ICCV*, 2019, pp. 1774–1783.

[34] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *CVPR*, 2017, pp. 7167–7176.

[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[36] P. S. R. Kishore, S. Das, P. S. Mukherjee, and U. Bhattacharya, "Cluenet: A deep framework for occluded pedestrian pose estimation," in *BMVC*, 2019.

[37] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *arXiv*, 2017.

[38] T. He, Z. Tian, W. Huang, C. Shen, Y. Qiao, and C. Sun, "An end-to-end textspotter with explicit alignment and attention," in *CVPR*, 2018, pp. 5020–5029.

[39] P. Lyu, M. Liao, C. Yao, W. Wu, and X. Bai, "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," in *ECCV*, 2018, pp. 67–83.

[40] Y. Sun, C. Zhang, Z. Huang, J. Liu, J. Han, and E. Ding, "Textnet: Irregular text reading from images with an end-to-end trainable network," in *ACCV*, 2018, pp. 83–99.

[41] N. Nayef, F. Yin, I. Bizid *et al.*, "ICDAR 2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt," in *ICDAR*, vol. 1, 2017, pp. 1454–1459.

[42] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *CVPR*, 2012, pp. 1083–1090.

[43] P. Lyu, C. Yao, W. Wu, S. Yan, and X. Bai, "Multi-oriented scene text detection via corner localization and region segmentation," in *CVPR*, 2018, pp. 7553–7563.

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

[45] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *CVPR*, 2017, pp. 1492–1500.

[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[47] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *ICML*, 2009, pp. 41–48.