W-A net: Leveraging Atrous and Deformable Convolutions for Efficient Text Detection

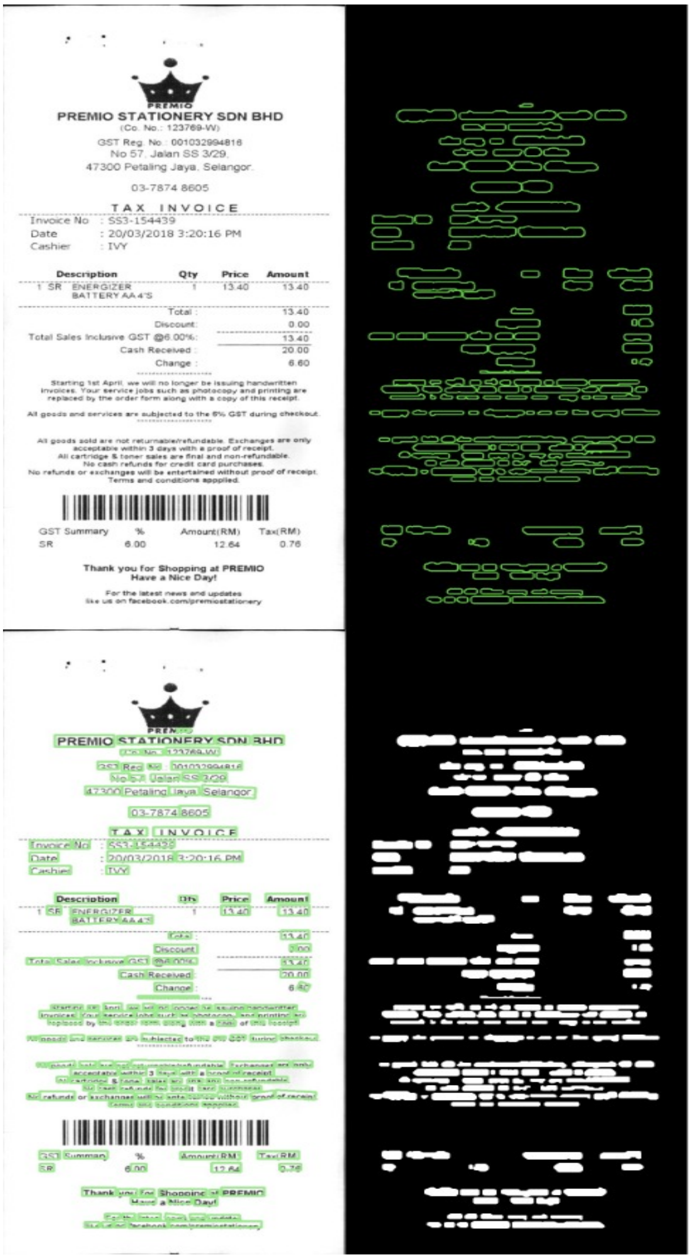
Sukhad Anand* and Zoya Khan†
Email: *sukhad_bt2k15@dtu.ac.in, †zoya1996k@gmail.com

***Abstract*—Scene text detection has been gaining a lots of focus in research. Even though the recent methods are able to detect text in complex background having complex shapes with a fairly good accuracy, they still suffer from issues of limited receptive field. These fail from detecting extremely short or long words hence failing in detecting text words precisely in document text images. We propose a new model which we call W-A net, because of it's W shape with the middle branch being Atrous convolutional layers. Our model predicts a segmentation map which divides the image into word and no word regions and also, a boundary map which helps to segregate closer words from each other. We use Atrous convolutions and Deformable convolutional layers to increase the receptive field which helps to detect long words in an image. We treat text detection problem as a single problem irrespective of the background, making our model suitable of detecting text in scene or document images. We present our findings on two scene text datasets and a receipt dataset. Our results show that our method performs better than recent scene text detection methods which perform poorly on document text images, especially receipt images with short words.**

I. INTRODUCTION

Information Extraction and Retrieval systems have received a lot of traction in the past few years due to the great importance of the task that they propose [1]. These systems aim at extracting valuable information from document images for aiding tasks like automated data entry, efficiently extracting user information from boarding passes in trains and airplanes, information extraction from forms and invoices [2], [3] and also aiding the design of products for assisting visually challenged users. A typical deep learning based Document Information Extraction pipeline involves a Text Detection system followed by a Text Recognition system and finally an Information Retrieval engine as described in [4].

In this paper, we focus on the first vital step of this pipeline, i.e. the task of text detection and localization in images, as state of the art text recognition systems have already attained impressive accuracy [5]. An efficient Text Localization technique play a crucial role in improving the accuracy of Text Recognition systems as it identifies the boundary boxes for text in the image which greatly assists Optical Character Recognition. While recent years have seen a lot of research being done in the domain of detecting text in scene images [6] and document images [3] in isolation, only very few models are robust enough to handle both scene text and document image scenarios.



PREMIO
PREMIO STATIONERY SDN BHD
(Co. No.: 123769-W)
GST Reg. No.: 001032994816
No 57, Jalan SS 3/29,
47300 Petaling Jaya, Selangor.
03-7874 8605

TAX INVOICE

Invoice No : SS3-154439
Date : 20/03/2018 3:20:16 PM
Cashier : IVY

Description	Qty	Price	Amount
1 SR ENERGIZER BATTERY AA 4'S	1	13.40	13.40

Total :	13.40
Discount:	0.00
Total Sales Inclusive GST @6.00%:	13.40
Cash Received :	20.00
Change :	6.60

Starting 1st April, we will no longer be issuing handwritten invoices. Your service jobs such as photocopy and printing are replaced by the order form along with a copy of this receipt.

All goods and services are subjected to the 6% GST during checkout.

All goods sold are not returnable/refundable. Exchanges are only acceptable within 3 days with a proof of receipt.
All cartridge & toner sales are final and non-refundable.
No cash refunds for credit card purchases.
No refunds or exchanges will be entertained without proof of receipt.
Terms and conditions applied.

GST Summary	%	Amount(RM)	Tax(RM)
SR	6.00	12.64	0.76

Thank you for Shopping at PREMIO
Have a Nice Day!

For the latest news and updates
like us on facebook.com/premiostationery

Fig. 1. This image shows the output of our proposed model on one of the image of SROIE dataset. The image depicts that our model is even able to read short words on a document and is easily able to segregate them from one another.

Also, while existing efficient lightweight text detection methods as proposed by Zhou et al. [6] perform fairly well on

[§] Equal contribution

scene text images, they suffer from two issues:

1) They are only able to detect text proportional to the receptive field of their model.
2) They are not able to detect vertical text instances easily due to lack of training data.

This limits their ability to detect long running text lines and vertical text instances in sign boards or even in logos of brands present in receipts or forms.

To handle these complexities, in this paper we propose a novel architecture of a W-A net for efficient text localization. We introduce Atrous Convolutions and Deformable Convolutions in our model to enhance the effective receptive field of the network, this makes our model robust enough to perform well with both document images and scene images. Our model is also agnostic to changes in the image geometries because of the presence of Deformable Convolutions.

Our contributions can be summarized as the following:

- A new network architecture which we call W-A net which consists of Atrous Convolution and Deformable Convolution layers which assist in capturing more detail due to the increased effective receptive field.
- We combine Atrous-II block features with the upsampled features. Atrous features are high resolution features and upsampled features from deeeper layers capture finer edge details. We combine both these features which significantly improves the accuracy.
- A robust model that efficiently handles text instances in all environments and orientations, i.e. in document and scene images both.
- We approach the problem of text detection as a segmentation problem where we divide the image into word and non word region predicting a binary map. We also predict a boundary map which helps to segment close words from one another.
- We present our findings on three different types of datasets with completely different nature - a scene text dataset (ICDAR 2015), a curved scene text dataset (CTW1500) and a receipt dataset (ICDAR 2019).

It can be seen that our approach performs better than the recent methods on scene text datasets and also on a receipt dataset. This proves that our model is insensitive to the length of words in an image.

II. Related work

Textual information retrieval has been a very important aspect in scene and document image understanding. An information retrieval pipeline involves text detection and recognition. Most of the cases involve pre-processing of image before actual text detection or recognition. The papers on scene text recognition focus on issues like curved text, irregular text or text distortions. They solve the issue of text recognition by first rectifying text and then recognising it. [7] use thin spline transformation for rectification. An attention-based mechanism is then used for recognition on rectified text which predicts letters of a word in a sequence. The model is end-to-end learnable. The parameters of thin-spline transformation are learned using spatial-transform network [8]. Linjie et al.[9] make use of deformable convolutions to recognize irregular text. Deformable convolutions help to recognize the text without rectification before recognizing it. Canjie et. al [10], combine deformable convolution kernels with recognition networks. They detect the offsets using the deformable kernels and then correct them before recognising text. Some methods use attention based mechanisms for recognition. A recent approach by Deli et al. [11], provides a different point of view by eliminating the use of RNN's for capturing attention and rather introducing a module to capture semantic reasoning. Christial et. al. [12], have proposed an architecture which does not have a rectification network which is an important part of most of the proposed approaches. They show how a simple model consisting of off shelf building blocks of neural network can be used directly for recognition without worrying about rectification.

As text recognition depends a lot on accuracy of text detection, if the bounding box does not enclose all the characters in a word, recognition can do nothing about it. This becomes a great issue in document image text retrieval, because accurate word detection is a great issue in document images, because, the text is focused on a small area as compared to scene text recognition. Text detection in document images also bring several complexities like distortions, occlusions, fading of text, deformations. Research on document text detection either focus on document restoration or simply text detection. Text detection in document images is often accomplished by detecting text lines. [13] have proposed approaches based on connected components which give significant results on document text extraction. [14] have proposed a corner detection based approach for text line and word detection for different languages. A recent approach by Manuel et al.[15] discusses an end-to-end handwritten text detection technique where they combine classification and regression to generate boxes around text words which are then fed into the recognition network.

Scene text detection has complexities like complex backgrounds, curved text and different illuminations at different points. These complexities have made scene text detection a more researched area than document text detection. Considering scene text detection is far more complex than document text detection, a document text detection can be modelled as scene text detection problem but with simpler backgrounds and other complexities as discussed above which do not occur in scenes. Scene text detection has been mostly modelled as a special case of object detection, hence the techniques for it are also derived many object detection problems.[16][17][18]. Even though these techniques give state of the art predictions, recognition accuracies with these techniques are still low, considering, most of the recognizers require tightly localised text which can't be detected by these methods due to arbitrary shapes of text which is not the case with object detection.

Recent scene text detection techniques fairly comprise of segmentation based, attention based or regression based

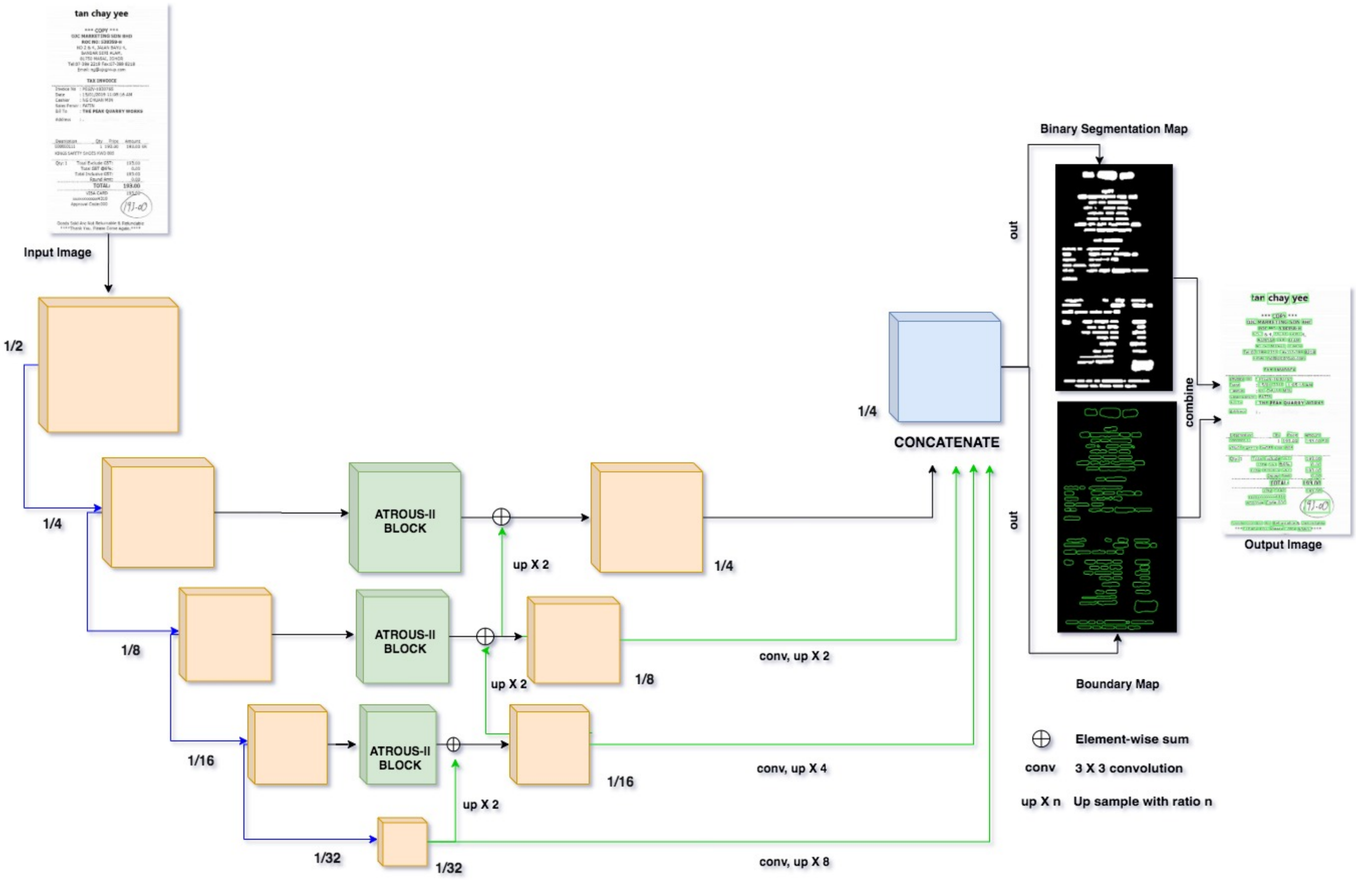


Fig. 2. Our model W-A net architecture consists of the input image undergoing a series of down-sampling operations and then upsampling with concatenation. The upsampled features are summed with the Atrous features to capture both finer and high resolution details.

methods and try to predict tight boundaries for text and even consider arbitrary shaped text, including curved text and perspective text. Most of these techniques use Resnet-52 or Resnet-18 as their backbone and fully convolution networks as their base architecture. Xinyu et al.[6] have proposed a regression-based technique which produces rotated and multi-oriented text boxes using FCN. Wenhao et al.[19] have also followed a similar strategy and produce rotated text boxes directly. TextBoxes++ [20] present an approach in which they detect arbitrary shaped text boxes using end-to-end trainable fully convolution networks and corrects them using regression. These techniques suffer from problem of low receptive fields which make them inappropriate for detection of long words and document word localisation. We have addressed this issue in our paper by increasing receptive field to capture long and arbitrary shaped text. Baoguang et al.[21] produce multi-oriented text boxes using segment linking where they produce small segments of words and text lines which are combined using the detected links to generate proposals.

Dena et al. [22] make use of fully convolution networks to segment document into text and no-text regions. Some other segmentation based methods tackle the problem of curved and irregularly shaped text. Wenhai et al. [23] propose a segmentation-based method where they segment text regions. They have used different scale of kernels to detect two words separately which are close to each other. TextSnake [24] make use of fully convolution network in an unconventional way. They represent text instances as sequence of ordered, overlapping disks centred at symmetric axes, each of which potentially variable radius and orientation. Minghui et al. [25] introduced differential binarization module to the segmentation network. They show how performance can be affected by introduction of this module.

III. Methodology

Previous methods on text localization consider document text detection or scene text detection as two different problems, and their solutions apply to only one of these. This is mostly because of the low receptive fields of the scene text detection networks and inability of document text detection frameworks to separate foreground from background. We contextualize text detection as a single problem and propose a network which generalises and performs effectively on both scene and document text detection. We propose a model which has a high receptive field, making it capable of detecting long words. Our technique involves binary segmentation, which is insensitive to shape and aids in detecting arbitrary shaped text in the images. Furthermore, our method helps to detect tightly localised bounding boxes which makes them suitable for textual recognition.

A. *Network Architecture*

Our network W-A net produces a segmentation map **S** and boundary map **B** which are then used to create bounding boxes around text words. We binarize the complete document and scene images into word and non word regions. The boundary map is necessary to separate and differentiate between two close words. Our network is a fully convolutional network (FCN) which uses Resnet-52[26] as its backbone. To increase the receptive field of our network, we introduce Atrous Convolution layers in the backbone network. We also employ Deformable Convolutions which provide sensitivity to arbitrary shapes of text words. We have designed our network in such a way that we are able to segment text words with any background and of any shape and are able to generate tight text boundaries. We apply Atrous-II blocks [27] parallel to 3rd, 4th and 5th stages of our backbone network. Instead of using one feature map to generate features of different scales and fusing them as done in ASPP module[28], we generate Atrous features at each stage to capture information at different resolution. These layers help to capture high resolution features of the image with a wider receptive field. U-net[29] has proved that information from multiple layers in the convolutional network helps to better estimate object boundaries. To capture both boundary details and wider receptive fields in a network, we combine the feature maps of the upsampled convolution features from lower layers which help to get finer details with the features from Atrous blocks at the mentioned convolution stages. The produced features are concatenated together to produce a rich feature map **F**. Then, feature **F** is used to predict both the segmentation map **S** and the boundary map **B**. The bounding boxes are generated from the segmentation map and the boundary map using simple formation module. We now discuss the various components of our network in detail.

B. *Receptive Field Enhancement*

One of the limitations of the architecture of most text detection techniques like [6] is not having a big enough receptive field to be able to effectively detect text in long sentences. To overcome this drawback our network uses Atrous-II blocks to maintain a uniform and strong receptive field to ensure detection of both short and long words. The effective receptive field of a network impacts how well the network makes use of the intricacies in the input image. Recent methods like [10] use various techniques to increase the receptive field of a network like stacking down-sampling layers with decreasing kernel size to increase receptive field or using sub-sampling to increase receptive field multiplicatively, cause the feature maps to become too small for pixel-level information to be retained. Another drawback of these methods is the large number of layers in the network and hence the increase in learnable parameters which arise. Recently, a lot of research has been done in exploiting high receptive field networks because of their efficient usage in tasks of medical image segmentation as described in [27]. Our model architecture focuses on enhancing the effective receptive field by employing both, Atrous-II blocks and Deformable convolutions on a ResNet-52 backbone.

*1) **Atrous Convolutions:*** Dilated Convolutions often known as Atrous Convolutions offer a unique way of enhancing the receptive field of a Deep Convolutional Neural Network (DCNN), all the while maintaining the same number of learnable parameters in the model. We employ Atrous-II blocks in our model as introduced by Zhou et al. [27], to help our model to effectively detect text instances. Using Atrous convolutions for increasing the receptive field comes at the cost of high memory usage, which is due to increased zeros, consequently propagating a higher dimensional feature map costs more; hence Atrous convolutions are often applied in tandem with convolutional layers to achieve a tradeoff of receptive field and memory. Atrous convolutions fare better by giving the same receptive field with lesser learnable parameters, no decrease in size of the feature map, removing the need of doing any upsampling techniques as there was no downsampling.

*2) **Deformable Convolutions:*** Deformable Convolutions [30] are very popular in computer vision because of their ability to provide a resilient receptive field and enhancing the ability of DCNNs to model geometric information because of the presence of leranable offsets in the kernel during training. Zhu et al. [31] explore the results of applying deformable convolutions to all layers of a DCNN which is not always feasible and desirable, in our DCNN we apply Deformable Convolutions in only the first two layers. These convolutions make our model robust even with the lack of training images of different geometries and aspect ratios.

C. *Boundary Detection*

The network outputs a Boundary Map which is used to generate the boundary of binary segmented text words. Boundary detection for separating words has also been discussed by Zhan et al. in [32]. The boundaries produced by their network is made using 4 segments. Instead of creating a boundary using segments, we propose a continuous approach. This allows us to detect boundaries of arbitrary shape and hence helps us in successfully segmenting text regions. We can then separate close words from each other and also help to accurately detect arbitrary shaped text. We generate groundtruth of our boundaries using the ground truth segmentation map. We propose boundary loss as the sum of product of softmax probability outputs of the boundary map and L2 distances between the predicted boundary and the groundtruth boundary map at a point on the predicted map.

$$L_t(\hat{y}, \hat{x}) = \frac{1}{n} \sum_{i \in B} \left\| (1 - \hat{p}_i)(\hat{y}_i - \hat{x}_i)^2 \right\| \tag{1}$$

D. *Label Generation*

Our network predicts a segmentation map and a boundary map. So, we need both these maps as input for training. In

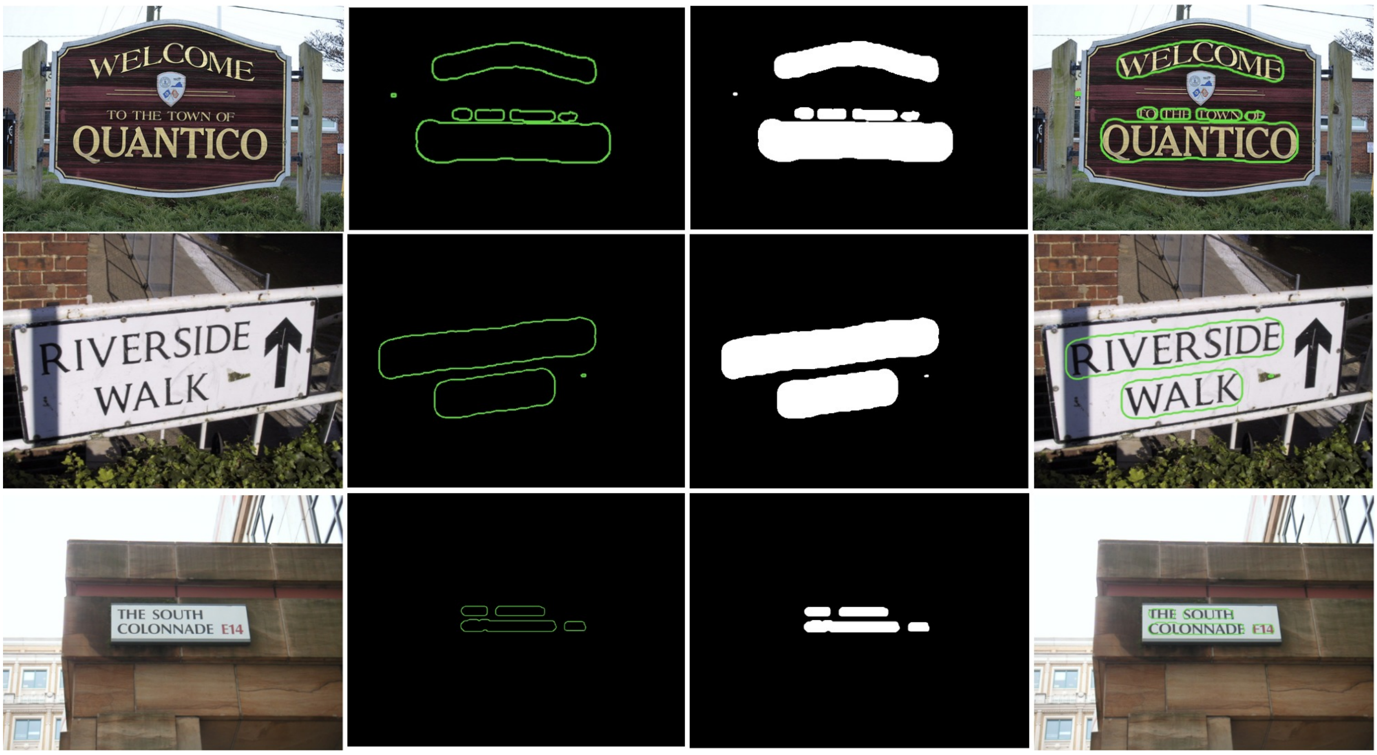


Fig. 3. This illustration depicts the results of our model on images from various datasets. We highlight the Segmentation Map, Boundary Map and final output prediction from the input image

most of the datasets we just have corners of the bounding boxes of words as labels. To generate labels, we have to convert an image into a binary map of word and non word regions and also produce boundaries of these segmented word regions. As we need tight boundaries, we cannot directly classify the complete region inside a bounding text box as a word and all remaining document as non-word region.

We follow the label generation technique as used by Liao et al. [25]. Each word is represented as a polygon with n vertices where n varies for different datasets. To generate tight bounding boxes we shrink the groundtruth polygon using Vatti clipping algorithm [33]. We mark the area inside this shrunk polygon as word region and all the other areas as non word region. Using this technique we binarize the complete image into word and non-word regions. This gives us a binary segmentation map S^i which we use for training.

To generate boundary map, we dilate the groundtruth polygon using the same offset which we used for shrinking using Vatti clipping algorithm. The gap between the shrunk polygon and the dilated polygon is used as groundtruth for the boundary map B^i.

We combine both the maps to generate the groundtruth labels $Y = \left\{S^{(i)}, B^{(i)}\right\}$, for each image i.

IV. Experiments

A. *Training*

Given a training set $T = \left\{X^{(i)}, Y^{(i)}\right\}$ where $X^{(i)}$ is the i-th image patch and $Y^{(i)}$ is the corresponding segmentation and boundary map, we train our proposed network. We pass the image through the proposed network to generate the boundary map and segmentation map. We calculate the loss as defined below and backpropagate to learn the network weights. We first pre-train our proposed network on SynthText[34] dataset for 50k iterations. We then fine tune our model on real-world datasets for 1000 epochs. In these experiments, we use Adam optimizer with a batch size of 64, momentum of 0.9 and a learning rate of 0.00001.All the experiments have been conducted on a workstation with 2.4 GHz CPU, 32 GB RAM, NVIDIA Tesla K80 GPU, running on Ubuntu 18.04.The data augmentation for the training data includes: Random Rotation, Random Cropping and Random Flipping. All the processed images are re-sized to 512X512.

*1) **Loss Function**:* For the evaluation of our network, we use a loss function that is a combination of the Binary Segmentation Map loss and a Boundary Map loss. The Binary Segmentation loss is further a combination of the Jaccard Index and the Binary Cross Entropy Loss. The usage of Jaccard Index, also known as an intersection over union evaluation metric is inspired by Iglovikov et al. in [35] as shown in the following equation:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cup B|} \tag{2}$$

As the Jaccard Index is not differentiable, we use the normal Jaccard Index as used by [36] this can be illustrated by:

$$J_m(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{i=N} \frac{y_i \hat{y}_i}{y_i + \hat{y}_i - y_i \hat{y}_i} \tag{3}$$

The binary cross entropy segmentation loss, where y and ŷ represent the two classes predicted at the pixel level for the image, can be summarized as :

$$L_s = -\frac{1}{N} \sum_{i=1}^{i=N} (y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \tag{4}$$

Therefore the final combined loss function of our network can be conceptualized in the following equation:

$$L = \beta L_t + \alpha L_s + (1 - \alpha)(1 - J) \tag{5}$$

We observe that the optimal value for α is 0.4 and β is 0.2.

B. Datasets

ICDAR 2015 dataset: This is the dataset provided by ICDAR 2015 Robust Reading Challenge on Incidental Scene Text Detection and recognition [37]. It consists of 1000 images for training and 500 images for testing. Ground truth comprises of boundary boxes for each word and their Unicode transcriptions.

ICDAR 2019 dataset: This is the dataset provided by ICDAR 2019 Robust Reading Challenge on Scanned Receipts OCR and Information Extraction (SROIE) and has 1000 scanned images. We focus on the Scanned Receipts Text Localization task. Each image consists of four main regions of text detection. Compared to previously available datasets for information extraction in receipts, this dataset is more robust as it contains blurred, poor resolution, folded receipts, long texts and small font sizes and other issues which make text detection even more challenging.

CTW1500 dataset: CTW1500[38] 2019 is a dataset which focuses on the curved text. It consists of 1000 training images and 500 testing images. The text instances are annotated in the text-line level.

C. Inference

We compare our proposed model with previous methods on three standard benchmarks, precision, recall and F1-Score. Each of the three datasets provide a different complexity. We evaluate our model on 3 different datasets with different complexities. We take each of scene text, curved text and document text dataset, to show how our network is able to correctly label text boxes in an image irrespective of the shape of text and the background. Table I compares our results on ICDAR 2015 scene text dataset. Our results show how our network is able to capture different backgrounds well and give better results than previous approaches. Previous approaches fail at detecting long words on scene text images. Introducing long receptive fields improves the text detection significantly.

Table II shows that our network is able to handle arbitrary shaped text in complex backgrounds. Introducing Atrous and Deformable convolutions help to tackle both the issues of receptive field and arbitrary shaped text instances.

TABLE I
COMPARISON OF PREVIOUS TEXT DETECTION METHODS ON ICDAR 2015 DATASET

Approach	Precision	Recall	F1-Score
CTPN[16]	74.2	51.6	60.9
EAST[6]	83.6	73.5	78.2
SSTD[39]	80.0	73.0	77.0
TextBoxes++[20]	87.2	76.7	81.7
PSENet[23]	86.9	**84.5**	85.7
TextSnake[24]	84.9	80.4	82.6
SegLink[21]	73.1	76.8	75.0
PixelLink[40]	85.5	82.0	83.7
SAE[41]	82.7	77.8	80.1
Ours	**89.34**	82.72	**85.90**

TABLE II
COMPARISON OF PREVIOUS TEXT DETECTION METHODS ON CTW1500 DATASET

Approach	Precision	Recall	F1-Score
CTPN[16]	60.4	53.8	56.9
EAST[6]	78.7	49.1	60.4
SegLink[21]	42.3	40.0	40.8
TextSnake[24]	67.9	85.3	75.6
TLOC[42]	77.4	69.8	73.4
PSENet[23]	**84.8**	79.7	82.2
SAE[41]	82.7	77.8	80.1
Ours	83.6	**86.1**	**84.83**

To illustrate our network's capability of detecting short text in simple backgrounds like document, we evaluate our network on the SROIE dataset and compare our results with previous approaches. We show how methods like TextSnake[cite], even though perform good on curved text instances, have a lower accuracy on simple text instances like documents. We show how our network is able to capture both the issues.

TABLE III
COMPARISON OF PREVIOUS TEXT DETECTION METHODS ON SROIE 2019 DATASET

Approach	Precision	Recall	F1-Score
CTPN[16]	79.22	61.6	68.9
EAST[6]	87.65	79.5	83.37
SSTD[39]	84.0	78.0	80.88
TextBoxes++[20]	89.2	79.7	84.18
PSENet[23]	89.5	82.5	85.85
TextSnake[24]	86.9	82.4	84.59
SegLink[21]	90.05	81.8	88.44
PixelLink[40]	91.5	85.3	88.29
SAE[41]	90.12	83.8	86.84
Ours	**92.7**	**86.8**	**89.65**

We show visualisation of our results on three datasets, highlighting the Segmentation Map output and the Boundary Map output along with the input image and predicted output.

V. Conclusion and Future Work

In this paper we highlight how increasing the effective receptive field of a DCNN aids in improving the segmentation capability by capturing high resolution features. Using Atrous features with up-sampled features helps to significantly improve the accuracy of our Segmentation Map and Boundary Map. The use of Deformable Convolutions helps to effectively capture the arbitrary shapes of text and also helps to detect different shapes of text words with complex backgrounds. We explore how introducing more context can help to detect even shorter words with very small size and capturing a very small area as shown in Fig 1. However, introducing Atrous and Deformable convolution layers comes at a high cost of memory and processing ability. This increases the computation time and renders our network inefficient for real time analysis. In future, we aim to reduce both processing and memory cost and make this a real time evaluation network.

References

[1] R. Gaizauskas and Y. Wilks, "Information extraction: Beyond document retrieval," *Journal of documentation*, vol. 54, no. 1, pp. 70–105, 1998.

[2] A. R. Katti, C. Reisswig, C. Guder, S. Brarda, S. Bickel, J. Höhne, and J. B. Faddoul, "Chargrid: Towards understanding 2d documents," *arXiv preprint arXiv:1809.08799*, 2018.

[3] X. Zhao, E. Niu, Z. Wu, and X. Wang, "Cutie: Learning to understand documents with convolutional universal text information extractor," *arXiv preprint arXiv:1903.12363*, 2019.

[4] P. Zhang, Y. Xu, Z. Cheng, S. Pu, J. Lu, L. Qiao, Y. Niu, and F. Wu, "Trie: End-to-end text reading and information extraction for document understanding," *arXiv preprint arXiv:2005.13118*, 2020.

[5] A. P. Tafti, A. Baghaie, M. Assefi, H. R. Arabnia, Z. Yu, and P. Peissig, "Ocr as a service: an experimental evaluation of google docs ocr, tesseract, abbyy finereader, and transym," in *International Symposium on Visual Computing*. Springer, 2016, pp. 735–746.

[6] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "East: an efficient and accurate scene text detector," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 5551–5560.

[7] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "Aster: An attentional scene text recognizer with flexible rectification." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2035–2048, 2019.

[8] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks." in *NIPS*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 2017–2025.

[9] Y. Gong, L. Deng, X. Lu, X. Yi, Z. Ma, and M. Xie, "Focus-enhanced scene text recognition with deformable convolutions." *CoRR*, vol. abs/1908.10998, 2019.

[10] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Advances in neural information processing systems*, 2016, pp. 4898–4906.

[11] D. Yu, X. Li, C. Zhang, J. Han, J. Liu, and E. Ding, "Towards accurate scene text recognition with semantic reasoning networks." *CoRR*, vol. abs/2003.12294, 2020.

[12] C. Bartz, J. Bethge, H. Yang, and C. Meinel, "Kiss: Keeping it simple for scene text recognition." *CoRR*, vol. abs/1911.08400, 2019.

[13] V. K. Koppula, A. Negi, and U. Garain, "Robust text line, word and character extraction from telugu document image." in *ICETET*. IEEE Computer Society, 2009, pp. 269–272.

[14] V. Yadav and N. Ragot, "Text extraction in document images: Highlight on using corner points." in *DAS*. IEEE Computer Society, 2016, pp. 281–286.

[15] M. Carbonell, J. Mas, M. Villegas, A. Fornés, and J. Lladós, "End-to-end handwritten text detection and transcription in full pages." in *WML@ICDAR*. IEEE, 2019, pp. 29–34.

[16] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network." in *ECCV (8)*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9912. Springer, 2016, pp. 56–72.

[17] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network," 2016, cite arxiv:1611.06779Comment: Accepted by AAAI2017.

[18] S. Zhang, Y. Liu, L. Jin, and C. Luo, "Feature enhancement network: A refined scene text detector." in *AAAI*, S. A. McIlraith and K. Q. Weinberger, Eds. AAAI Press, 2018, pp. 2612–2619.

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

[20] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "Textboxes: A fast text detector with a single deep neural network," 2016, cite arxiv:1611.06779Comment: Accepted by AAAI2017.

[21] B. Shi, X. Bai, and S. J. Belongie, "Detecting oriented text in natural images by linking segments." in *CVPR*. IEEE Computer Society, 2017, pp. 3482–3490.

[22] D. Bazazian, R. Gomez, A. Nicolaou, L. G. i Bigorda, D. Karatzas, and A. D. Bagdanov, "Improving text proposals for scene images with fully convolutional networks." *CoRR*, vol. abs/1702.05089, 2017.

[23] X. Li, W. Wang, W. Hou, R.-Z. Liu, T. Lu, and J. Yang, "Shape robust text detection with progressive scale expansion network." *CoRR*, vol. abs/1806.02559, 2018.

[24] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, "Textsnake: A flexible representation for detecting text of arbitrary shapes." *CoRR*, vol. abs/1807.01544, 2018. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1807.htmlabs-1807-01544

[25] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-time scene text detection with differentiable binarization." *CoRR*, vol. abs/1911.08947, 2019.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, cite arxiv:1512.03385Comment: Tech report.

[27] X.-Y. Zhou, J.-Q. Zheng, P. Li, and G.-Z. Yang, "Acnn: a full resolution dcnn for medical image segmentation," *arXiv preprint arXiv:1901.09203*, 2019.

[28] C. Mayer, R. Timofte, and G. Paul, "Towards closing the gap in weakly supervised semantic segmentation with dcnns: Combining local and global models," 2018, cite arxiv:1808.01625.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[30] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[31] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9308–9316.

[32] C. Xue, S. Lu, and F. Zhan, "Accurate scene text detection through border semantics awareness and bootstrapping," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11220. Springer, 2018, pp. 370–387.

[33] B. R. Vatti, "A generic solution to polygon clipping," *Commun. ACM*, vol. 35, no. 7, pp. 56–63, 1992.

[34] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition." *CoRR*, vol. abs/1406.2227, 2014. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1406.htmlJaderbergSVZ14

[35] V. Iglovikov, S. S. Seferbekov, A. Buslaev, and A. Shvets, "Ternausnetv2: Fully convolutional network for instance segmentation." in *CVPR Workshops*, vol. 233, 2018, p. 237.

[36] V. Iglovikov, S. Mushinskiy, and V. Osin, "Satellite imagery feature detection using deep convolutional neural network: A kaggle competition," *arXiv preprint arXiv:1706.06169*, 2017.

[37] 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 *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2015, pp. 1156–1160.

[38] Y. Liu, L. Jin, S. Zhang, and S. Zhang, "Detecting curve text in the wild: New dataset and new solution," *CoRR*, vol. abs/1712.02170, 2017.

[39] W. Zhu, J. Lou, Q. Xia, and M. Ren, "Single shot text detector with rotational prior boxes," *Neural Processing Letters*, vol. 49, no. 3, pp. 863–877, 2019.

[40] D. Deng, H. Liu, X. Li, and D. Cai, “Pixellink: Detecting scene text via instance segmentation,” in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, S. A. McIlraith and K. Q. Weinberger, Eds. AAAI Press, 2018, pp. 6773–6780.

[41] S. Long, X. He, and C. Yao, “Scene text detection and recognition: The deep learning era,” *CoRR*, vol. abs/1811.04256, 2018.

[42] Z. Liu, G. Lin, S. Yang, F. Liu, W. Lin, and W. L. Goh, “Towards robust curve text detection with conditional spatial expansion,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 7269–7278.