End-to-End Scene Text Spotting at Character Level

Zobeir Raisi
Vision and Image Processing Lab, University of Waterloo, ON, N2L 3G1, Canada
John Zelek
Vision and Image Processing Lab, University of Waterloo, ON, N2L 3G1, Canada
Email: {zraisi, jzelek}@uwaterloo.ca

Abstract

This work utilizes the new object detection framework, namely Detection using Transformers (DETR), to spot the characters in the wild images, which offers simpler and robust end-to-end architecture than the previous methods. The proposed framework leverages an adaptive feature extraction to better focus on the position of character regions and a bounding box loss function that is more precise in spotting characters with different scales and aspect ratios.

To evaluate our proposed architecture's effect, we conduct experiments on the ICDAR benchmark designed explicitly for character-level text detection, namely the ICDAR13 dataset. Experimental results show that the proposed method outperforms the state-of-the-art detectors when tested on the mentioned benchmark.

1 Introduction

Reading text from wild images is one of the challenging problems in the computer vision community due to many variations in text appearance, sizes, shapes, aspect ratios, font styles, perspective distortion, and the complicated image background. For reading, a text in a scene requires two stages: locate the text and then recognize the character in the detected regions, which are called scene text detection and scene text recognition. Some methods combine these two stages, which leads to an end-to-end detection and recognition (scene text spotting).

Inspired by deep-learning frameworks like Convolutional Neural Network (CNN) [1, 2] and Recurrent Neural Network (RNN) [3], many end-to-end scene text detection and recognition methods [4–8] proposed. Some of these methods achieved superior performance of text spotting at word-level in different benchmark datasets [9–11]. However, these CNN and RNN based methods require several handcrafted components such as anchor generation, non-maximum suppression (NMS) in *regression-base* methods, or multiple processing stages (*e.g.* label generation) in *segmentation-based* method to detect following by a rectification module before to output the sequences of characters using RNN. Furthermore, Some of these models, as described in [12, 13] show poor performance when characters in the text are vertical or partially occluded.

As mentioned above, previous scene text spotting approaches aim to output word instances whose primary components are characters. Therefore, we aim to design a simple and end-to-end framework that directly and precisely extracts the characters from the given image and then combines the extracted characters to form the final word. To achieve this goal, we utilize state-of-the-art Transformer-based techniques that alleviate the issues of previous CNN-based methods.

Transformer [14] is an attention-based pipeline that, after achieving superior performance in sequence modelling and machine translation tasks [15], recently emerged in many computer vision fields and achieved state-of-the-art results in many benchmarks [16, 17]. Current state-of-the-art object detectors [18–22] mainly inspired by self-attention mechanism in Transformers outperformed prior Convolution Neural Networks (CNN) models [23]. For example, Detection using Transformer (DETR) [18], was the first encoder-decoder Transformer-based detector proposed a new concept for object detection framework. DETR uses a new technique called object queries and task object detection as a set prediction problem [23]. In contrast to other detectors, it removed the need to design hand-designed components like anchor design and non-maximum suppression (NMS) post-processing and directly detects objects in the given image using so-called object queries. However, DETR has low accuracy on small objects and slow convergence during training [19, 23].

Many recent works proposed efforts to alleviate the issues mentioned for [18], for example, Deformable-DETR [19] aims to design data-dependent sparse attention to address the small object detection problem of [18] and achieved higher precision performance and fewer training epochs. Pyramid Vision Transformer (PVT) [20] is a hierarchical pure Transformer backbone that achieved superior performance in classification, object detection and segmentation tasks. In order to decline the sequence length in the given input and preserve the channel dimensions fixed, PVT utilizes a non-overlapping patch partition followed by a linear patch embedding, respectively. This backbone can be used accompanied by a Transformer framework like [18] to predict dense objects efficiently.

Sparse R-CNN [21] proposed a sparse algorithm for object detection without relying upon dense candidate regions. In order to detect objects, it first generates a random sparse set of boxes and then iteratively performs classifications and detection of the candidate boxes. In a recent work, Deformable Patch-based Transformer (DPT) [22] presented DePatch that adaptively split images in a data-driven way which address the problem of PVT [20] that uses the predefined fixed-patched. DePatch forces the network to concentrate on desired object regions and extract more semantic formations in patches with different positions and scales. DPT achieved state-of-the-art performance on image classification and object detection.

In this work, we only focus on character spotting by leveraging the DTER [18] as our baseline detector. The contribution of these works are: (1) We propose a new Transformer based model based on [18] by modifying its feature extraction backbone and prediction head by leveraging a robust bounding box loss function. (2) We compare state-of-the-art transformer-based methods on spotting the characters of the wild images with our proposed architecture. (3) We provide quantitative and qualitative results to show the performance of our proposed model.

2 Methodology

Figure 1 shows the proposed architecture. The framework of our proposed method mostly follows the encoder-decoder detector form [18]. The network first adaptively extracts image features using a DPT-Small [22] backbone from different small patches; The resulting feature set is passed to a transformer encoder. For decoding, a fixed set of learned embeddings called object queries are passed through a transformer decoder. The feature vectors tests obtained are fed to shared fully connected layers that directly predict each query's class and bounding box set. The Bipartite matching loss is used for training the network, which leverages the Hungarian matching algorithm [25] for comparing and establishing a one-to-one mapping between N queries and N ground-truths [18]. The prediction head outputs rectangular bounding boxes $b = [x, y, w, h]^\top$ can encase the character region by simplifying defining (x, y) as the bounding box's center point coordinates, and w, h representing the box's width and height respectively. To train the network, we also modify the prediction head, along with the loss and matching functions as described in below.

Loss function: The bounding box loss function of [18] uses a linear combination of ℓ_1 and GIoU loss. Let $\hat{b}_i$ and b_j denote the i^{th} predicted and j^{th} ground truth bounding boxes, respectively, then we define our loss function as:

$$\mathcal{L}_{\text{box}}(\hat{b}_i, b_j) = \lambda_1 \mathcal{L}_{reg}(\hat{b}_i, b_j) + \lambda_2 \mathcal{L}_{\alpha-\text{GIoU}}(\hat{b}_i, b_j) \tag{1}$$

where λ_1 and $\lambda_2 \in \mathbb{R}$ are hyper-parameters, and $\mathcal{L}_{reg}(\cdot)$ and $\mathcal{L}_{\alpha-\text{GIoU}}(\cdot)$ are the rectangular bounding box loss functions based on regression and $\alpha-$GIoU. The $\alpha-$GIoU is defined as [26]:

$$\mathcal{L}_{\alpha\text{-GIoU}} = 1 - IoU^\alpha + \left(\frac{|C \setminus (B \cup B^{gt})|}{|C|}\right)^\alpha, \tag{2}$$

where $\mathcal{L}_{\alpha\text{-IoU}} = 1 - IoU^\alpha$, C denotes the smallest convex shape enclosing b_i and $\hat{b}_i$. In our experiments, the $\alpha = 3$ showed better performance.

For regression, we use the Smooth-ln based Regression Loss as in [24]. The regression loss is then defined as:

$$\mathcal{L}_{reg}(\hat{b}_i, b_j) = (|\hat{b}_i - b_j| + 1)\ln(|\hat{b}_i - b_j| + 1) - |\hat{b}_i - b_j| \tag{3}$$

where $|\cdot|$ demonstrates the absolute operator.

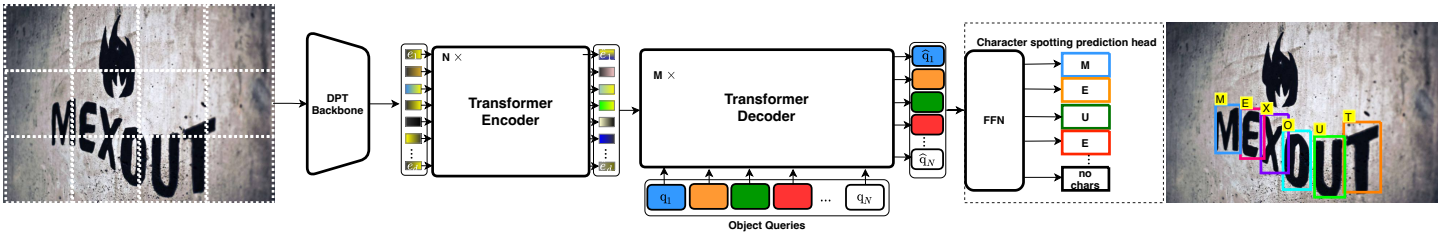


Fig. 1: The proposed end-to-end character-level text spotting framework, modified from [24].

3 Experimental Results

In this section, we first compare our proposed method with state-of-the-art Transformer-based detectors and then present some qualitative results to show the model's performance. Finally, we provide an ablation study to investigate the effect of the different components in the proposed pipeline.

Implementation details: We adopt the DETR [19] architecture as our main framework with a DPT-small [2] backbone for feature extraction. The number of object queries are set to 300 and AdamW [27] optimizer is used to optimize the parameters of the model. We use horizontal flip and and resize the images similar to [18] for augmentation. We first pre-train our proposed model and methods in comparison on $100k$ images of Synth-text [28] with character level annotations for 8 epochs and then fine-tuned on the ICDAR13 dataset to ensure the training converges. We train our model with a batch size of 2 per GPU using 4 Tesla V100 GPUs and a learning rate (LR) of 1×10^{-4}.

Datasets: The ICDAR13 dataset [29] is the only benchmarks that includes both word-level and character-level annotations using rectangular boxes containing 229 and 233 images for training and testing. Most of the text instances of this dataset are horizontal and high-resolution. Since ICDAR13 is the only well-known benchmark dataset that contains character-level annotations, we conduct our experiment on this dataset. However, we provide some qualitative sample results in section 3.2 on other arbitrary-shape text datasets including Total-Text [10] and CTW-1500 [11] to better show the performance of our model.

Evaluation metric: To our best knowledge, this is the first work that focuses on character spotting; there is no evaluation metric to measure the performances of the predicted characters in the scene text detection community. Nevertheless, we can task characters as different classes of objects; Thus, we can use mean average precision (AP) as our evaluation metric adopted as a standard in many recent object detection algorithms to spot 36 alphanumerical (10 digits + 26 capital) characters directly in the images.

3.1 Quantitative Results

To evaluate the performance of the proposed method, We compare it with DETR [18], PVT [20], Sparce R-CNN [21], and DPT [22]. The quantitative comparison is shown in Table 1. Our proposed method outperformed the state-of-the-art detectors by a large margin,$\sim 4\%$ compare to the best detector in AP performance. It also performed better in the spotting of small, medium, and large characters. The baseline DETR [18] not only performed poorly on small and large characters, but it also required more training epochs to converge on the ICDAR13 dataset. On the other hand, with a lower number of training iterations, PVT significantly outperforms DETR by $\sim 8\%$. While Sparce-RCNN outperformed the PVT in overall AP by $\sim 2\%$ in reading better of medium and large characters, it showed poor performance in spotting small characters. In contrast, DPT performed better in small character spotting and achieved the second-best performance in terms of AP.

3.2 Qualitative Results

Figure 3 shows the qualitative results on some challenging sample images. As seen, the proposed model is robust in spotting small, medium, large and even complex fonts characters compared to the baseline model. It also performed well on spotting of partially occluded and oriented characters as shown in Figure 3(b) and Figure 3(c), respectively.

Fig. 2: Qualitative results of the proposed method in out of distribution samples from Total-text [10] and CTW-1500 [11] datasets. The proposed method detects characters in arbitrary-shape text instances.

To see the generalization ability of the proposed method, we also provided some qualitative result of arbitrary-shape text of Total-text [10] and CTW-1500 [11] datasets, where model was agnostic to the text instance of them. As shown in Figure 2 the model was able to detect and recognize precisely the characters in various text instances of the given images.

Ablation study: To assess the added value of the various components in our model, we performed an extensive ablation study on ICDAR13 datasets. Table 2 summarizes the obtained results.

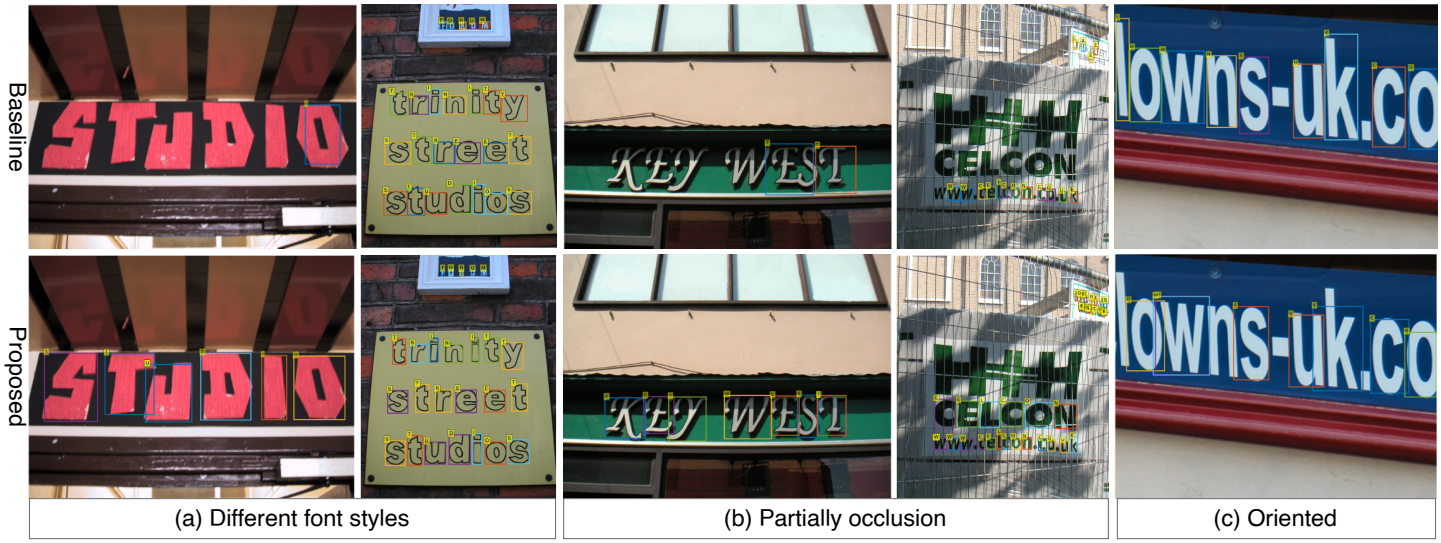
We started the experiments by baseline model that uses a ResNet-50 backbone for feature extraction, GIoU+ℓ_1 loss for bounding box regression; the model achieved an AP performance of 0.41. We then replaced the backbone with PVT-small yielding an AP=0.46, which outperformed the baseline; We found that using DPT-small as backbone led to further performance boost compared to PVT-small backbone. We finally replaced the baseline bounding box losses with $\alpha-$GIoU+Smooth-ln loss and achieved the best performance on the mentioned dataset by improving $\sim 4\%$.

4 Conclusion

This paper has leveraged a new end-to-end Transformer-based architecture for character spotting in the wild images. The proposed method has leveraged Deformable-Patch (DPT) as a feature extraction backbone and a bounding box loss function for reading characters with different sizes, scales, and aspect ratios in the wild images. We experimented with ICDAR benchmark dataset to compare our proposed method's performance with that of the recent state-of-the-art object detection approaches. Experimental results have shown that the proposed method outperforms the state-of-the-art methods, including recent Transformer based detectors, in terms of mean average precision. Our end-to-end robust character level detector is an essential step towards the word or text-line detection, which remains part of our future work. As future work, we are interested in addressing the occluded text challenge and geometric distortions by improving the proposed method's scheme.

Table 1: Comparing the character spotting performance of our proposed methods with state-of-the-art detectors [18, 20–22] on ICDAR13 [29] dataset.The best performance is highlighted in **bold**.

Model-Name	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	epochs
DETR [18]	0.49	0.78	0.57	0.48	0.58	0.41	700
PVT [20]	0.57	0.83	0.68	0.55	0.67	0.53	200
Sparse R-CNN [21]	0.59	0.80	0.70	0.49	0.69	0.65	200
DPT [22]	0.62	0.86	0.76	0.61	0.68	0.58	200
Proposed	**0.66**	**0.89**	**0.78**	**0.64**	**0.72**	**0.63**	**200**



(a) Different font styles (b) Partially occlusion (c) Oriented

Fig. 3: Qualitative comparison of the baseline [18] and proposed methods on some of the challenging images of ICDAR13 dataset. Best viewed when zoomed.

Table 2: Ablation study of our model using different components. The models trained only on the train set of ICDAR13 and no synthetic images used for pre-training. The best performance is shown in **bold**.

model	backbone	bounding box loss	AP
Baseline	ResNet50	GIoU+ℓ_1	0.41
Baseline-2	PVT-Small	GIoU+ℓ_1	0.46
Baseline-3	DPT-Small	GIoU+ℓ_1	0.48
Proposed	DPT-Small	$\alpha-$GIoU+Smooth-ln	**0.52**

Acknowledgments

We would like to thank the Ontario Centres of Excellence (OCE), the Natural Sciences and Engineering Research Council of Canada (NSERC), and ATS Automation Tooling Systems Inc., Cambridge, ON, Canada for supporting this research work.

References

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

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Conf. on Comp. Vision and Pattern Recognit. (CVPR)*, pp. 770–778, 2015.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[4] M. Busta, L. Neumann, and J. Matas, "Deep textspotter: An end-to-end trainable scene text localization and recognition framework," in *Proc. IEEE Int. Conf. on Comp. Vision*, 2017, pp. 2204–2212.

[5] 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 *Proc. Eur. Conf. on Comp. Vision (ECCV)*, 2018, pp. 67–83.

[6] W. Liu, C. Chen, and K.-Y. K. Wong, "Char-net: A character-aware neural network for distorted scene text recognition," in *Proc. AAAI Conf. on Artif. Intell.*, 2018.

[7] Y. Liu, H. Chen, C. Shen, T. He, L. Jin, and L. Wang, "Abcnet: Real-time scene text spotting with adaptive bezier-curve network," in *Proc. IEEE/CVF Conf. on Comput. Vision and Pattern Recognit.*, 2020, pp. 9809–9818.

[8] A. Singh, G. Pang, M. Toh, J. Huang, W. Galuba, and T. Hassner, "Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text," in *Proc. IEEE/CVF Confon. Comput. Vision and Pattern Recognit.*, 2021, pp. 8802–8812.

[9] 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 *Proc. Int. Conf. on Document Anal. and Recognition (ICDAR)*, 2015, pp. 1156–1160.

[10] C. K. Ch'ng and C. S. Chan, "Total-text: A comprehensive dataset for scene text detection and recognition," in *Proc. IAPR Int. Conf. on Document Anal. and Recognit. (ICDAR)*, vol. 1, 2017, pp. 935–942.

[11] L. Yuliang, J. Lianwen, Z. Shuaitao, and Z. Sheng, "Detecting curve text in the wild: New dataset and new solution," in *arXiv preprint arXiv:1712.02170*, 2017.

[12] Z. Raisi, M. A. Naiel, P. Fieguth, S. Wardell, and J. Zelek, "Text detection and recognition in the wild: A review," *arXiv preprint arXiv:2006.04305*, 2020.

[13] Z. Raisi, M. A. Naiel, P. Fieguth, S. Wardell, and J. Zelek, "Challenges of deep learning-based text detection in the wild," *Journal of Computational Vision and Imaging Systems*, vol. 6, no. 1, pp. 1–5, 2021.

[14] 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.

[15] S. Chaudhari, G. Polatkan, R. Ramanath, and V. Mithal, "An attentive survey of attention models," *arXiv preprint arXiv:1904.02874*, 2019.

[16] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *arXiv preprint arXiv:2101.01169*, 2021.

[17] Z. Raisi, M. A. Naiel, G. Younes, S. Wardell, and J. Zelek, "2lspe: 2d learnable sinusoidal positional encoding using transformer for scene text recognition," in *18th Conference on Robots and Vision (CRV)*, 2021, pp. 119–126.

[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *arXiv preprint arXiv:2005.12872*, 2020.

[19] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[20] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *arXiv preprint arXiv:2102.12122*, 2021.

[21] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang *et al.*, "Sparse r-cnn: End-to-end object detection with learnable proposals," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 454–14 463.

[22] Z. Chen, Y. Zhu, C. Zhao, G. Hu, W. Zeng, J. Wang, and M. Tang, "DPT: Deformable patch-based transformer for visual recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2899–2907.

[23] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A survey of visual transformers," *arXiv preprint arXiv:2111.06091*, 2021.

[24] Z. Raisi, M. A. Naiel, G. Younes, S. Wardell, and J. S. Zelek, "Transformer-based text detection in the wild," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021, pp. 3162–3171.

[25] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[26] J. He, S. M. Erfani, X. Ma, J. Bailey, Y. Chi, and X.-S. Hua, "Alpha-iou: A family of power intersection over union losses for bounding box regression," in *Proc. Neural Information Processing Systems*, 2021.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[28] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proc. IEEE Conf. on Comp. Vision and Pattern Recognit.*, 2016, pp. 2315–2324.

[29] 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 *Proc. Int. Conf. on Document Anal. and Recognition*, 2013, pp. 1484–1493.