PBFORMER: CAPTURING COMPLEX SCENE TEXT SHAPE WITH POLYNOMIAL BAND TRANSFORMER

Anonymous authors
Paper under double-blind review

1 APPENDIX

1.1 RELATIONS TO TESTR AND FEWBETTER

PBFormer has multiple efficient and useful designs compared with TESTR (Zhang et al., 2022) and FewBetter (Tang et al., 2022).

Compared with TESTR:

- PBFormer is single-stage without relying on intermediate bounding box results. In contrast, TESTR adopts the two-stage pipeline. It predicts the bounding box in each feature point of the transformer encoder's output, then selects topK boxes based on confidence to embed them into positional embeddings and reference points, as Fig. 1 (a)'s red block shows.
- PBFormer adopts a coarse-to-fine strategy to generate reference points for deformable attention modules. The first decoder's output encodes 2-d residuals to move the reference point into a better localization. In contrast, TESTR uses the same reference points for all deformable attention modules. Fig. 1 (a) and (c)'s orange arrows illustrate the differences.
- PBFormer only utilizes two layers of transformer encoders and decoders. In comparison, TESTR uses six layers of transformer encoders and decoders.
- PBFormer is only pre-trained on CurvedSynthText, while TESTR is pre-trained on a combination of three datasets, *i.e.*, CurvedSynthText, MLT2017, and Total-Text.

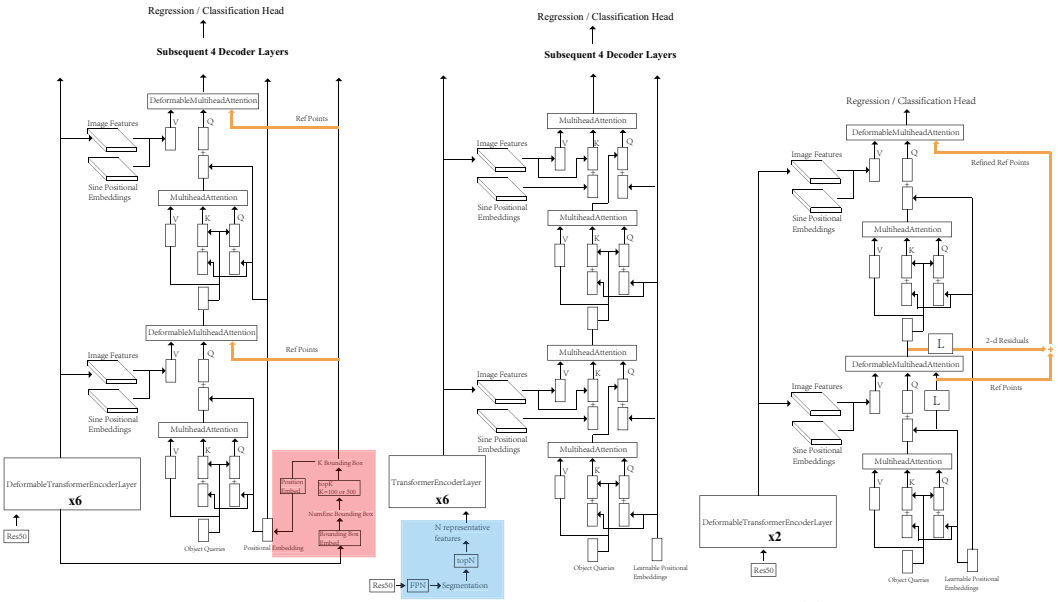
Compared with FewBetter:

- PBFormer does not need FPN in the backbone or generate segmentation maps. Differently, FewBetter has an FPN network to generate convincing segmentation masks for feature selections, as Fig. 1 (b)'s blue block shows.
- PBformer utilizes a deformable transformer while FewBetter uses the original transformer. Besides, we also use two transformer encoder and decoder layers, while FewBetter uses the six for encoder and decoder, respectively.

1.2 EFFICIENCY'S COMPARISONS OF DIFFERENT TEXT REPRESENTATION

The polynomial band is more efficient than segmentation masks and polygon points during inference, as Fig. 2 shows. The polynomial band is more efficient than Bezier points during ground truth generations.

- Polygon points need spline interpolation because the number of network's output is too sparse for evaluation, such as 16 points in TESTR, while more points are needed to compute accurate IoUs in MMOCR's evaluation protocol (Kuang et al., 2021), especially for curved texts. Therefore, as Fig. 2 shows, after dividing points into splines, each of which is fitted by a polynomial, then evenly sampling. In contrast, PB avoids the above two-stage iterative procedure because PB direct contains four curves for the whole contour, which generates a dense contour by sampling.
- Ground truths of Bezier points are calculated by the **least square algorithm** based on original polygonal annotations (Liu et al., 2020), which consumes additional time and resources for generating ground truth for extra data. Differently, PB directly utilizes the polygonal annotations without additional consumption to transform raw annotations.



(a) TESTR (b) FewBetter (c) PBFormer

Figure 1: **Detailed Network Comparisons with TESTR and FewBetter.** (a) NumEnc denotes the sequence length of encoder's output.

1.3 Ground Truth of Top, Bottom, Left and Right Generation.

In common text datasets, a text instance is annotated by $2K$ discrete points. For example, $K = 5$ in Total-Text and $K = 7$ in CTW1500. More importantly, these points are annotated along with human reading hobbies. They are ordered in counterclockwise order, and the first point is always the top-left of the first character. The former K points $\mathbf{p}_1, ...\mathbf{p}_K$ and the latter K points $\mathbf{p}_{K+1}, ...\mathbf{p}_{2K}$ are located as Fig. 3 shows (green for the former K, blue for the latter K).

Therefore, the generating process of ground truths for the top, bottom, left, and right curve follows subsequent five steps:

1. Divide $2K$ annotations into four Sets. $\mathbf{S}^1 = \{\mathbf{p}_1, \ldots, \mathbf{p}_K\}$, $\mathbf{S}^2 = \{\mathbf{p}_{K+1}, \ldots, \mathbf{p}_{2K}\}$, $\mathbf{S}^3 = \{\mathbf{p}_K, \mathbf{p}_{K+1}\}$, and $\mathbf{S}^4 = \{\mathbf{p}_{2K}, \mathbf{p}_1\}$.
2. Assuming $\mathbf{S}^1$ and $\mathbf{S}^2$ would be fitted by a mapping $f : x \to y$, judge whether assumed functions are both single-valued. If yes, go to step 3. If not, go to step 4.
3. $\mathbf{S}^1$ and $\mathbf{S}^2$ both use the form $y = f(x)$; $\mathbf{S}^3$ and $\mathbf{S}^4$ both use the form $x = f(y)$. If $\mathbf{S}^1$'s middle point is higher than $\mathbf{S}^2$'s, $\mathbf{S}^1$ becomes the ground truth for the top curve, and $\mathbf{S}^2$ becomes the ground truth for the bottom curve, and vice versa. If $\mathbf{S}^3$'s middle point is more left than $\mathbf{S}^4$'s, $\mathbf{S}^3$ becomes the ground truth for the left curve, and $\mathbf{S}^4$ becomes the ground truth for the right curve, and vice versa. Go to step 5.
4. $\mathbf{S}^1$ and $\mathbf{S}^2$ both use the form $x = f(y)$; $\mathbf{S}^3$ and $\mathbf{S}^4$ both use the form $y = f(x)$. If $\mathbf{S}^1$'s middle point is more left than $\mathbf{S}^2$'s, $\mathbf{S}^1$ becomes the ground truth for the left curve, and $\mathbf{S}^2$ becomes the ground truth for the right curve, and vice versa. If $\mathbf{S}^3$'s middle point is higher than $\mathbf{S}^4$'s, $\mathbf{S}^3$ becomes the ground truth for the top curve, and $\mathbf{S}^4$ becomes the ground truth for the bottom curve, and vice versa. Go to step 5.
5. Sample $\mathbf{S}^1$, $\mathbf{S}^2$, $\mathbf{S}^3$, and $\mathbf{S}^4$ evenly to generate dense supervision points.

Single-valued condition. A mapping $f : x \to y$ which is used to fit a set of order points satisfies the single-valued condition if and only if:

$$x_1 < x_2 < ... < x_n \vee x_1 > x_2 > ... > x_n, \tag{1}$$

where $x_1, \ldots, x_n$ is the x-coordinate sequence of order points, and n denotes the number of points.

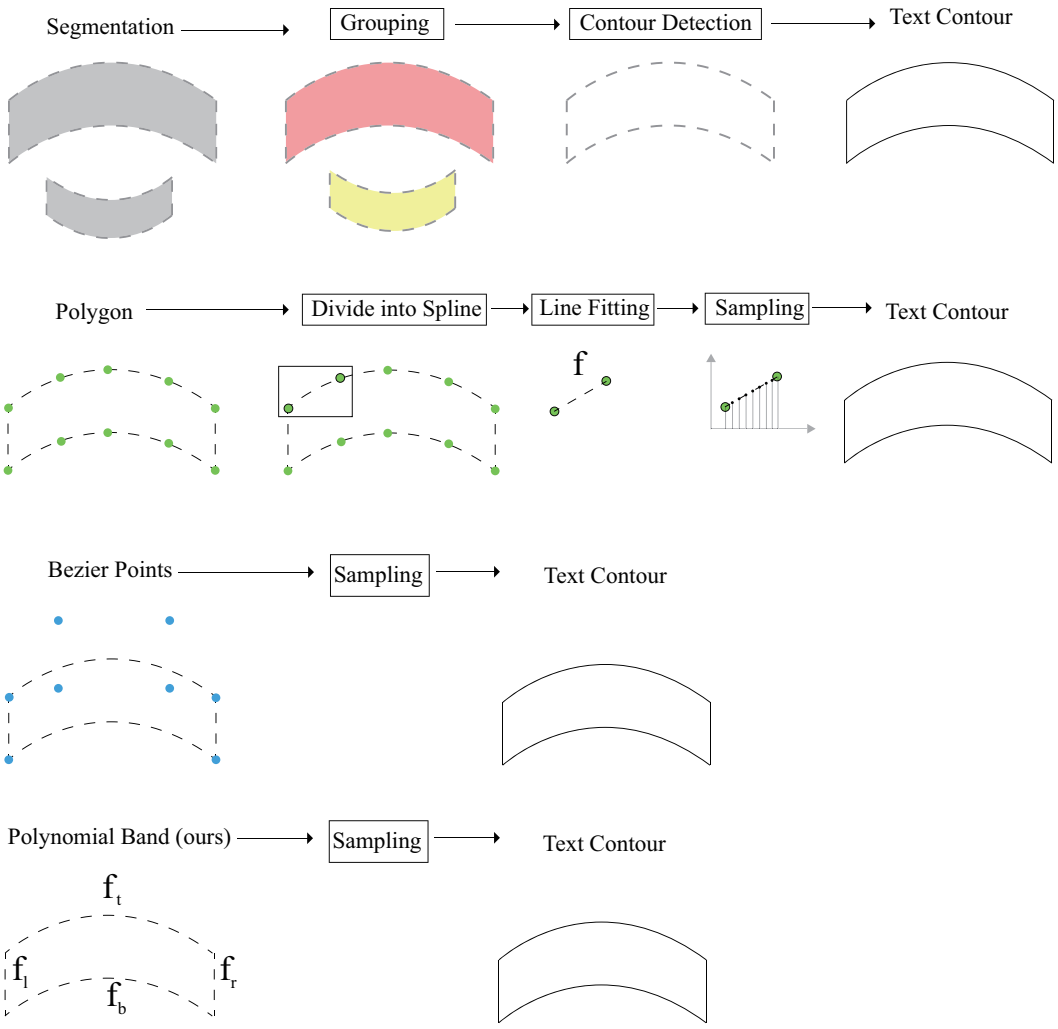


Figure 2: **Comparisons of the process from network's output representation to final text contour.** For segmentation maps, we only show main post-processing since there might be other practical processes such as outlier removal, distance (or directional) maps refinement, etc.

As Fig. 3 shows, we illustrate the processes for a horizontal reading direction text "COMPANY" and a vertical reading direction text "DISTILLERY."

1.4 More Visualizations on attention maps.

In Fig. 4, we select some representative attention maps of the cross-scale pixel attention module. The CPA is capable of attending to text regions at a suitable layer adaptively. We see CPA learned to attend at a single layer if texts have similar sizes. When different texts have size various, CPA learned to attend to relatively small texts at the swallow layer while large texts at the deep layer.

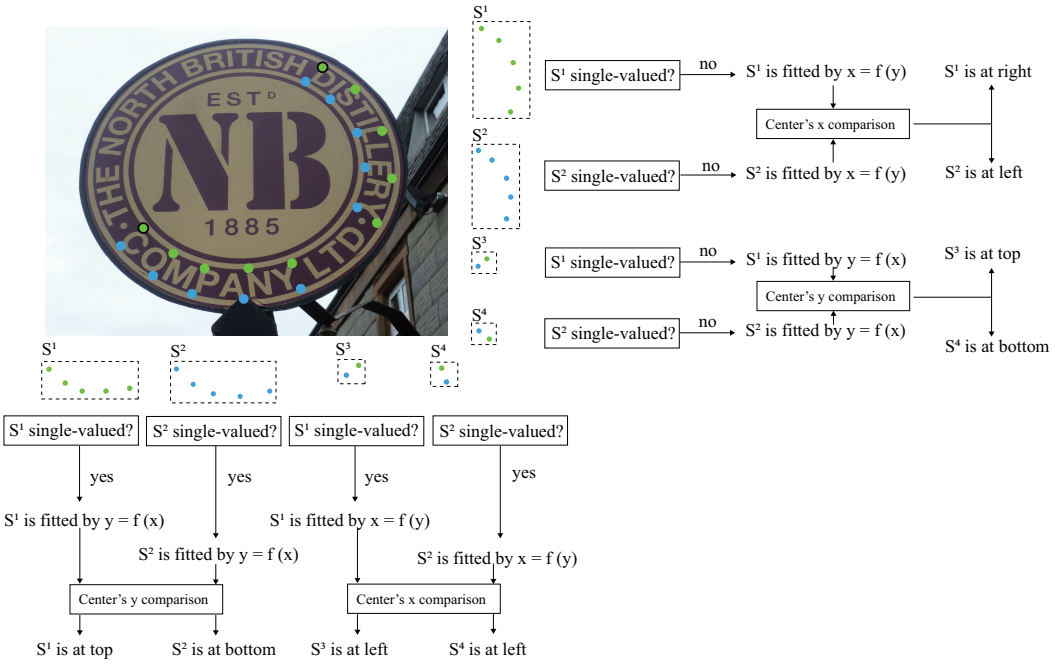


Figure 3: **Visualization of generating ground truth for the top, bottom, left, and right sides from raw annotations.** The black rounded green dots represent the first annotated point for text instances.

1.5 Failure cases.

In Fig. 5, we visualize failure cases in two situations: (1) texts which are extremely small as Fig. 5(a) shows; (2) texts which occupy a minority in common datasets, such as chinese-lingual texts in Fig. 5(b) and texts processed into symbols or beautified by art-style.

The limitations may inspire our future work. Firstly, enlarging the input's resolution can improve the detection of small texts, but it further consumes more computations. Secondly, collecting multi-lingual texts or adopting art-style augmentation on the common texts for training will improve the recall of the results.

1.6 More qualitative results.

In Fig. 6 and Fig. 7, we present more qualitative results on CTW1500 and Total-Text. PBFormer predicts not only accurate contour for curved texts but also handles multi-oriented data successfully. More importantly, when texts are crowded, PBFormer also distinguishes them clearly. Results indicate that our PBFormer is robust in various scenes.

References

Zhanghui Kuang, Hongbin Sun, Zhizhong Li, Xiaoyu Yue, Tsui Hin Lin, Jianyong Chen, Huaqiang Wei, Yiqin Zhu, Tong Gao, Wenwei Zhang, Kai Chen, Wayne Zhang, and Dahua Lin. Mmocr: A comprehensive toolbox for text detection, recognition and understanding. *arXiv preprint arXiv:2108.06543*, 2021.

Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In *CVPR*, pp. 9806–9815. Computer Vision Foundation / IEEE, 2020.

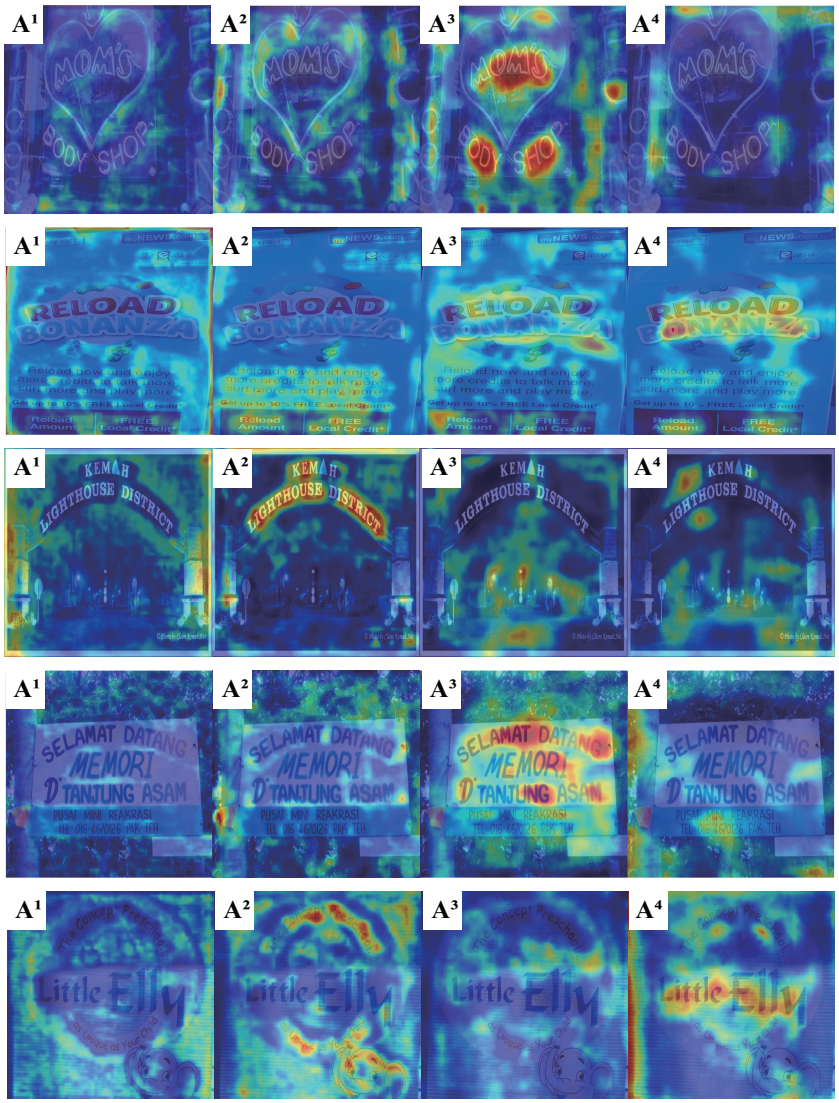


Figure 4: **CPA's attention maps on Total-Text.** $\mathbf{A}^1$, $\mathbf{A}^2$, $\mathbf{A}^3$ and $\mathbf{A}^4$ represent the attention map weighting the multi-scale feature from network's swallow to deep layers.

(a) (b) (c) (d) (e)

Figure 5: **Visualization of failure cases.** The yellow dashed boxes denote the failed detections.

Jingqun Tang, Wenqing Zhang, Hongye Liu, Mingkun Yang, Bo Jiang, Guanglong Hu, and Xiang Bai. Few could be better than all: Feature sampling and grouping for scene text detection. In *CVPR*. Computer Vision Foundation / IEEE, 2022.

Xiang Zhang, Yongwen Su, Subarna Tripathi, and Zhuowen Tu. Text spotting transformers. In *CVPR*. Computer Vision Foundation / IEEE, 2022.

Curved Texts

Crowded Texts

Multi-oriented Texts

Figure 6: **More Qualitative results on CTW1500.** We demonstrate different types of texts, such as curved, crowded, or multi-oriented.

Curved Texts

Crowded Texts

Multi-oriented Texts

Figure 7: **More Qualitative results on Total-Text.** We demonstrate different types of texts, such as curved, crowded, or multi-oriented.