Toward Understanding WordArt: Corner-Guided Transformer for Scene Text Recognition # Supplementary File

Xudong Xie[1], Ling Fu[1], Zhifei Zhang[2], Zhaowen Wang[2], and Xiang Bai[1](✉)

[1] Huazhong University of Science and Technology, China
{xdxie,ling_fu,xbai}@hust.edu.cn
[2] Adobe Research, USA
{zzhang,zhawang}@adobe.com

1 More Examples from the WordArt Dataset

Fig. 1 demonstrates more artistic text images from our proposed WordArt Dataset. It is a very challenging problem to accurately recognize these artistic texts because of the diverse fonts, extreme deformation, and word effects.

Fig. 1. Artistic text examples from the WordArt dataset

✉ Corresponding author

2 More Visualization of Corner Points

As illustrated in Fig. 2, corner points provide a robust representation for the artistic text image, suppressing the interference of appearance and deformation. The corner points of a character are almost invariant and stable. For blurred and low-contrast images, corners can still capture the most critical locations.

Fig. 2. Visualization of corner point detection. Corner points (blue dots) indicate the most critical locations in images that contain rich visual information

3 More Visualization of the Encoder Feature Map

Fig. 3 shows the feature map of the final output from our corner-guided encoder. Assisted by our corner-query cross-attention mechanism, the encoder can accurately focus on the position of each character and even the strokes.

4 More Qualitative Recognition Results

Fig. 4 shows some qualitative recognition results on WordArt, SVTP [3], ICDAR 2015 [2], and CUTE80 [4]. Our CornerTransformer can cope with artistic texts containing complex fonts, ligatures, and overlaps. It can also recognize extremely curved and deformed texts. Besides, gradient-based corner detection is robust to image resolution, noise, and blur, so CornerTransformer performs well on SVTP and ICDAR 2015, achieving state-of-the-art results.

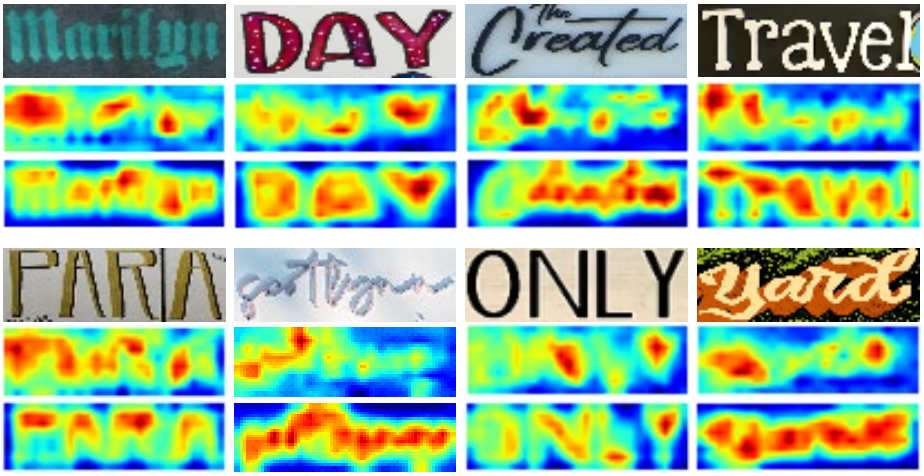
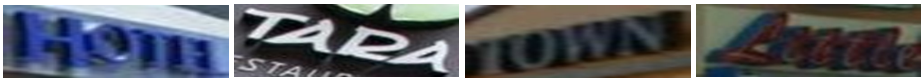


Fig. 3. Visualization for the feature map of the encoder output. First row: input images; Second row: feature maps of the baseline; Third row: feature maps of the baseline equipped with the corner-query cross-attention

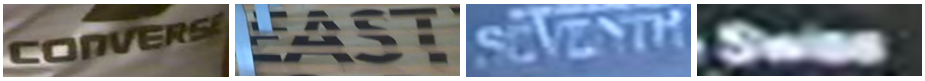
jody gady lady | mi55 mi55 miss | proxim prozim prosim | negs might night

(a) Qualitative results on WordArt



how hote hotel | ta12a tai2a tara | jonn iown town | date lun little

(b) Qualitative results on SVTP [3]



convers convers! converse | last fast east | sevent sci.crypt seventh | senee suns swiss

(c) Qualitative results on ICDAR 2015 [2]

2otbas hotahs motors | lundekvan lundekvain lundekvam | ctrrbucks strrbucks starbucks | tall able table

(d) Qualitative results on CUTE80 [4]

Fig. 4. Qualitative recognition results on WordArt and irregular benchmarks. Each example is along with the results from ABINet-LV [1], our baseline and the proposed CornerTransformer. Hard examples are successfully recognized by CornerTransformer

References

1. Fang, S., Xie, H., Wang, Y., Mao, Z., Zhang, Y.: Read like humans: autonomous, bidirectional and iterative language modeling for scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7098–7107 (2021) 3
2. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., et al.: Icdar 2015 competition on robust reading. In: 2015 13th international conference on document analysis and recognition (ICDAR). pp. 1156–1160. IEEE (2015) 2, 3
3. Phan, T.Q., Shivakumara, P., Tian, S., Tan, C.L.: Recognizing text with perspective distortion in natural scenes. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 569–576 (2013) 2, 3
4. Risnumawan, A., Shivakumara, P., Chan, C.S., Tan, C.L.: A robust arbitrary text detection system for natural scene images. Expert Systems with Applications **41**(18), 8027–8048 (2014) 2, 3