# Turning a CLIP Model into a Scene Text Detector (Supplementary Material)

## 1. Appendix

### 1.1. Datasets

**ICDAR2013** [5] is high-resolution English dataset for focused scene text detection, including 229 images for training and 233 images for testing.

**ICDAR2015** [4] is a multi-oriented text detection dataset for English text that includes 1,000 training images and 500 testing images. Scene text images in this dataset were taken by Google Glasses without taking care of positioning, image quality, and viewpoint.

**MSRA-TD500** [12] is a multi-language dataset that includes English and Chinese, including 300 training images and 200 testing images. We also include extra 400 training images from HUST-TR400 [11] following the previous methods [6, 13].

**CTW1500** [7] consists of 1,000 training images and 500 testing images which focuses on the curved text. The text instances are annotated in the text-line level by polygons with 14 vertices.

**Total-Text** [2] contains 1,255 training images and 300 testing images. The text instances are labeled at the word level. It includes horizontal, multi-oriented, and curved text shapes.

**ArT** [1] includes 5,603 training images and 4,563 testing images. It is a large-scale multi-lingual arbitrary-shape scene text detection dataset. The text regions are annotated by the polygons with an adaptive number of key points. Note that it contains Total-Text and CTW1500.

**MLT17** [9] includes 9 languages text representing 6 different scripts annotated by quadrangle. It has 7,200 training images, 1,800 validation images, and 9,000 testing images. We use both the training set and the validation set in the finetune period.

**MLT19** [8] is a large-scale multi-lingual scene text detection datasets. It contains 10,000 training images and 10,000 testing images, and labeled at word level.

**SynthText** [3] It contains 800k synthetic images generated by blending natural images with artificial text, which are all word-level annotated.

**TextOCR** [10] is a large-scale high quality scene text datasets collected from Open Images[1]. It contains 30 words on average per image. It has 24,902 training images and 3,232 testing images, and is annotated with polygons.

### 1.2. More Quantitative Results

**Multi-lingual Real-to-real Adaptation.** We conducted multi-lingual generalization ability experiments as shown in Table 1. The results show that the pluggable TCM can also benefit to multi-lingual scenarios text detection via leveraging the pretrained knowledge of CLIP, which demonstrates the effectiveness of our method for domain adaptation.

| Method | MLT17 → MLT19 |
|---|---|
| DBNet [6] | 47.4 |
| TCM-DBNet | **67.5** |

Table 1. Real-to-real adaptation. F-measure (%) is reported.

**Ablation Study for the Different Predefined Language Prompt.** We conducted ablation study on the predefined language prompt with different string using TCM-DBNet in Table 2. Results show that without predefined language prompt the performance is harmed. In addition, it can be seen that there is little performance variation with different predefined language prompt.

| Predefined language prompt | IC15 |
|---|---|
| "Text" | 89.2 |
| "A set of arbitrary-shape text instances" | 89.0 |
| "The pixels of many arbitrary-shape text instances" | 88.9 |
| without predefined language prompt | 87.7 |

Table 2. Ablation study of the different predefined language prompt.

**Ablation Study for Training with Large-scale Dataset.** We conducted ablation study of training TCM-DBNet on IC15 with extra TextOCR [10] data. As shown in Table 3, when using additional large-scale TextOCR as training data, our model can achieve further improvement, suggesting the compatibility of our method with large-scale datasets.

**Ablation study for CLIP Backbone Generalization.** We conducted ablation study to investigate the generalization performance of DBNet by directly replacing the backbone of

[1] Open Images Link

| Model | Training data | F (%) |
|---|---|---|
| TCM-DBNet | IC15 | 89.2 |
| TCM-DBNet | IC15+TextOCR | 90.4 |

Table 3. Ablation study of training TCM-DBNet on IC15 with extra TextOCR data.

DBNet with CLIP backbone, as shown in Table 4. It shows that the CLIP-R50 can indeed bring benefit for generalization. However, integrating with TCM, the performance can be significantly improved. It suggests that directly using the pre-trained CLIP-R50 is not strong enough to improve the generalization performance of the existing text detector, which further indicates that synergistic interaction between the detector and the CLIP is important.

| Model | Backbone | ST → IC13 | ST → IC13 |
|---|---|---|---|
| DBNet | R50 | 71.7 | 64.0 |
| DBNet | CLIP-R50 | 73.1 | 67.4 |
| TCM-DBNet | CLIP-R50 | 79.6 | 76.7 |

Table 4. Ablation study on CLIP backbone. R50 means ResNet50.

## 1.3. More Visualization Results

**Conditional Cue.** We visualize the t-SNE of the generated conditional cue (**cc**) on six datasets, as illustrated in Fig. 1. The structured distribution indicates our model has learned the distribution of every domain dataset in high-dimensional feature space, which is useful for improving the generalization ability.

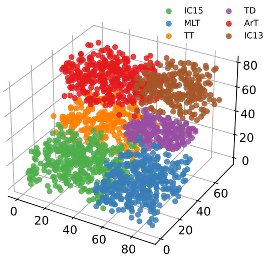

Figure 1. t-SNE of conditional cue (**cc**). MLT short for MLT17.

**Visual Prompt.** Fig. 2 - Fig. 5 are more qualitative results of the image embedding $\boldsymbol{I}$ and the generated visual prompt $\tilde{\boldsymbol{I}}$ on CTW1500, Total-Text, MSRA-TD500, and ICDAR2015, respectively. The visual prompt $\tilde{\boldsymbol{I}}$ has contains fine-grained information of text regions.

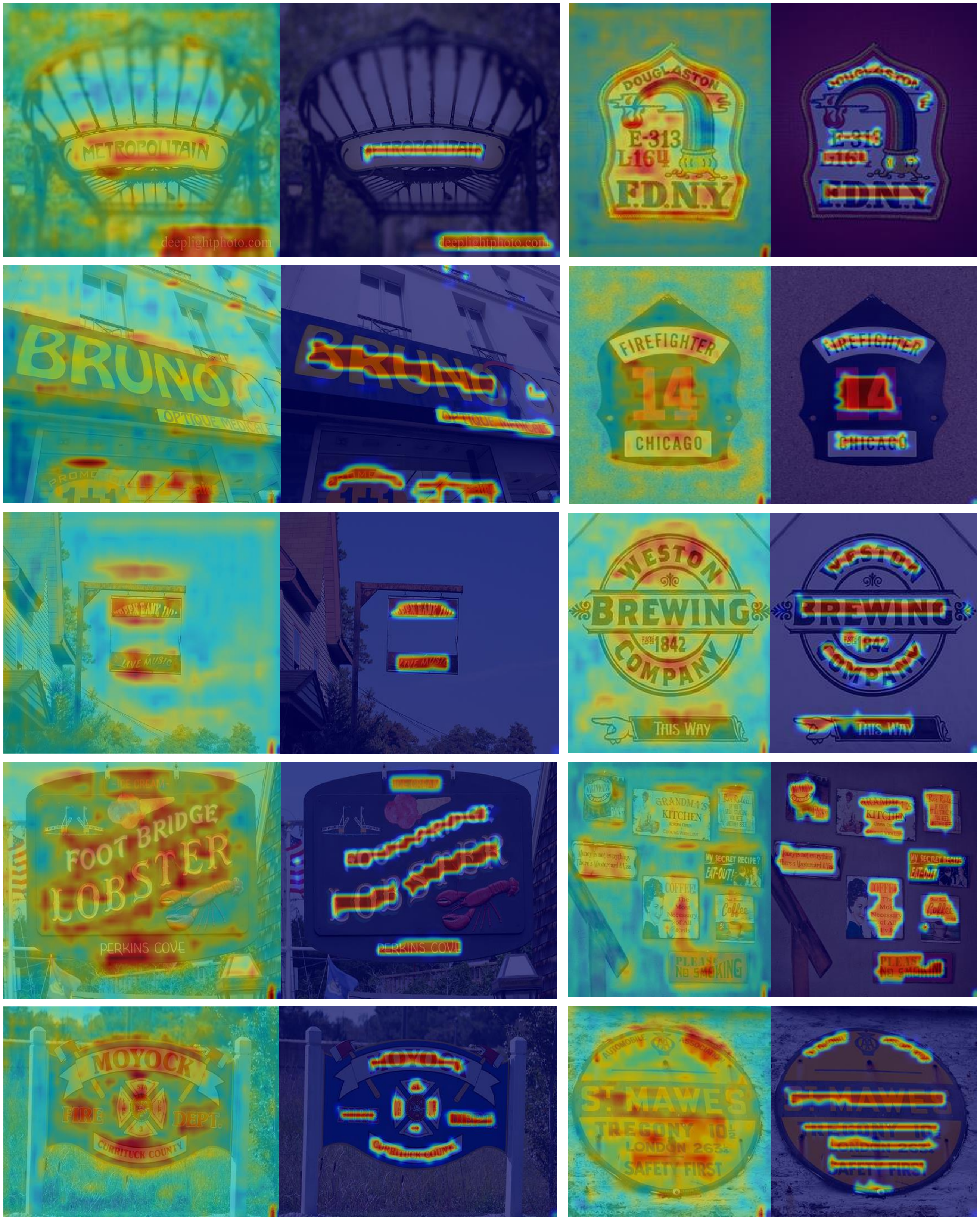

Figure 2. Visualization results of our method on CTW1500. For each pair, the left is the image embedding $\boldsymbol{I}$, and the right is the generated visual prompt $\tilde{\boldsymbol{I}}$. Best view in screen.

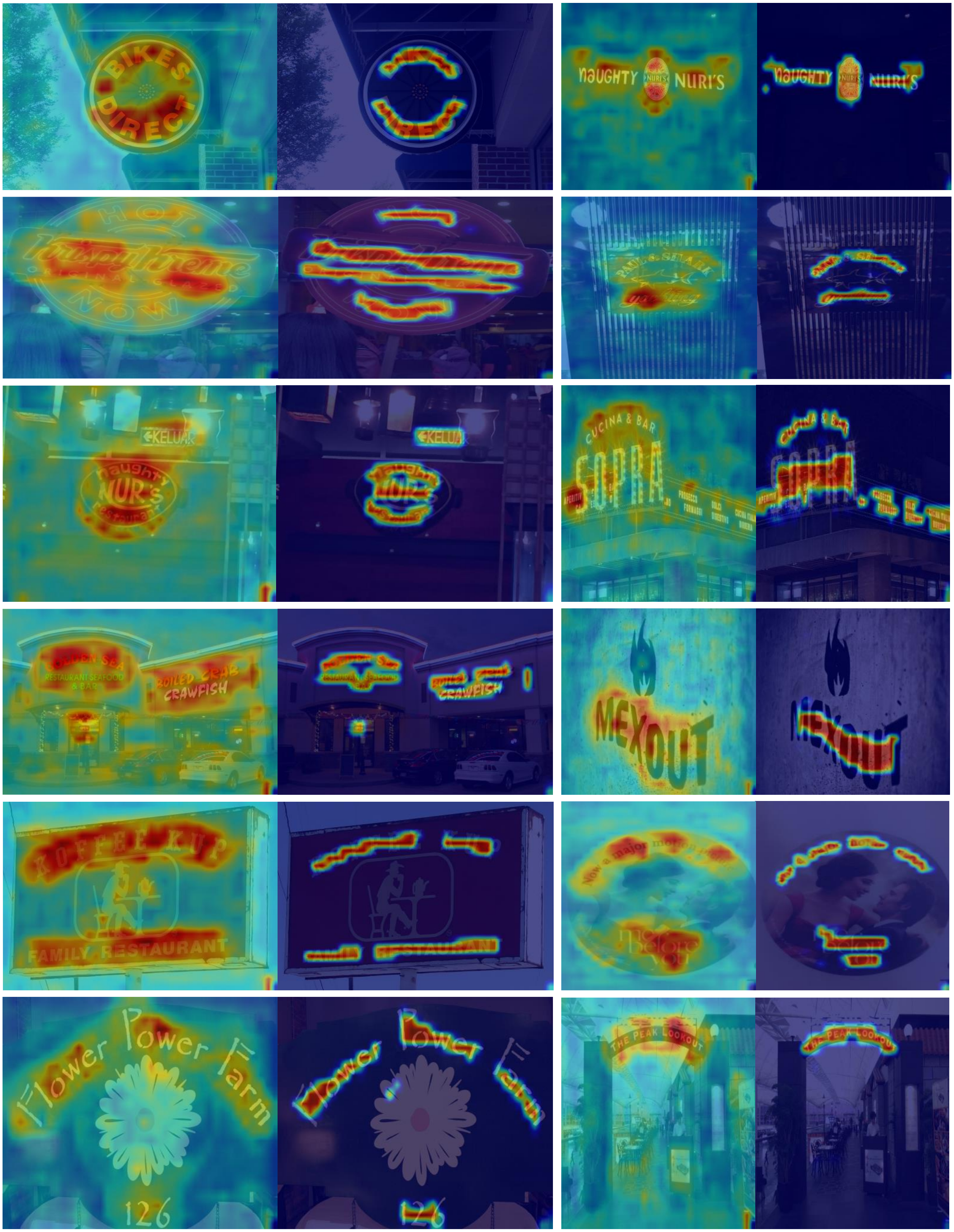

Figure 3. Visualization results of our method on Total-Text. For each pair, the left is the image embedding $\boldsymbol{I}$, and the right is the generated visual prompt $\tilde{\boldsymbol{I}}$. Best view in screen.

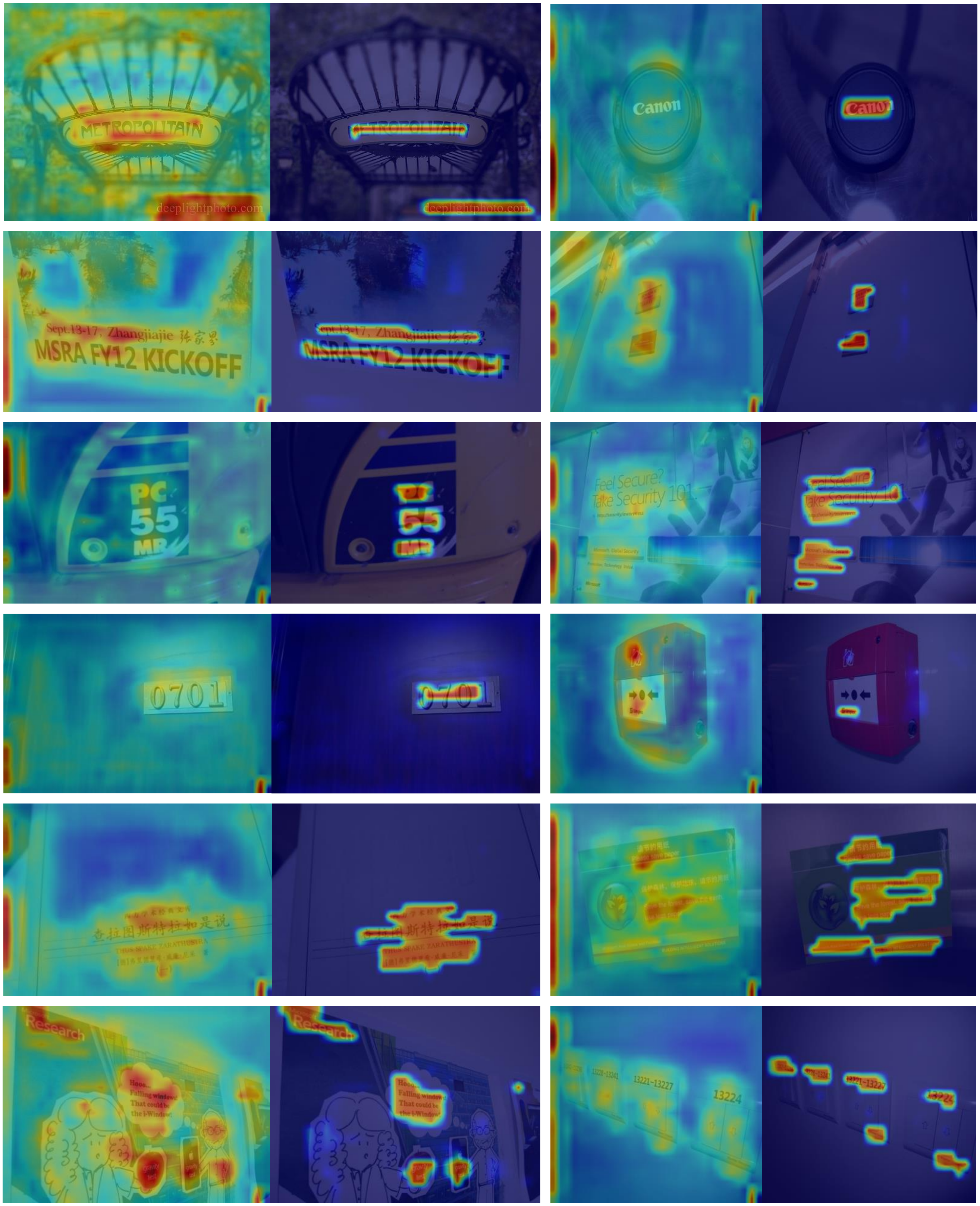

Figure 4. Visualization results of our method on MSAR-TD500. For each pair, the left is the image embedding $\boldsymbol{I}$, and the right is the generated visual prompt $\tilde{\boldsymbol{I}}$. Best view in screen.

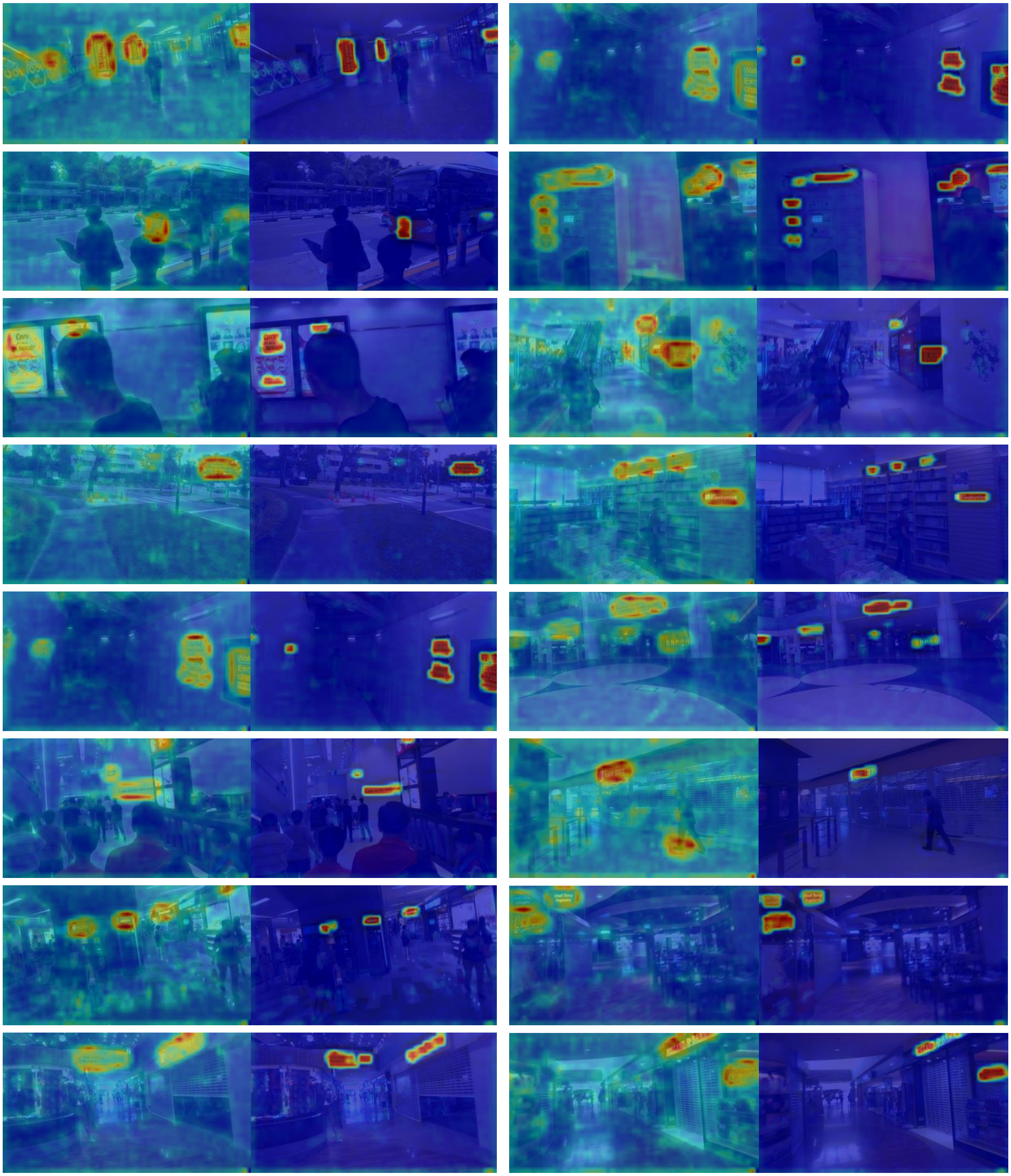

Figure 5. Visualization results of our method on ICDAR2015. For each pair, the left is the image embedding $\boldsymbol{I}$, and the right is the generated visual prompt $\tilde{\boldsymbol{I}}$. Best view in screen.

# References

[1] Chee-Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. ICDAR2019 Robust Reading Challenge on Arbitrary-Shaped Text (RRC-ArT). In *ICDAR*, pages 1571–1576, 2019. 1

[2] Chee-Kheng Ch'ng, Chee Seng Chan, and Cheng-Lin Liu. Total-text: toward orientation robustness in scene text detection. *IJDAR*, pages 1–22, 2019. 1

[3] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2315–2324, 2016. 1

[4] D. Karatzas, L. Gomez-Bigorda, et al. ICDAR 2015 competition on robust reading. In *ICDAR*, pages 1156–1160, 2015. 1

[5] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, M. Iwamura, Lluís Gómez i Bigorda, Sergi Robles Mestre, Joan Mas Romeu, David Fernández Mota, Jon Almazán, and Lluís-Pere de las Heras. Icdar 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1484–1493, 2013. 1

[6] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *AAAI*, pages 11474–11481, 2020. 1

[7] Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and Sheng Zhang. Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recognition*, 90:337–345, 2019. 1

[8] Nibal Nayef, Yash Patel, Michal Busta, Pinaki Nath Chowdhury, Dimosthenis Karatzas, Wafa Khlif, Jiri Matas, Umapada Pal, Jean-Christophe Burie, Cheng-lin Liu, et al. ICDAR2019 Robust Reading Challenge on Multi-lingual Scene Text Detection and Recognition–RRC-MLT-2019. In *ICDAR*, pages 1454–1459, 2019. 1

[9] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, Wafa Khlif, Muhammad Muzzamil Luqman, Jean-Christophe Burie, Cheng-Lin Liu, and Jean-Marc Ogier. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification - rrc-mlt. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 1454–1459, 2017. 1

[10] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8798–8808, 2021. 1

[11] Cong Yao, Xiang Bai, and Wenyu Liu. A unified framework for multi-oriented text detection and recognition. *IEEE Transactions on Image Processing*, 23:4737–4749, 2014. 1

[12] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Detecting texts of arbitrary orientations in natural images. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1083–1090. IEEE, 2012. 1

[13] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: An efficient and accurate scene text detector. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2642–2651, 2017. 1