# Primitive Representation Learning for Scene Text Recognition Supplementary Material

Ruijie Yan Liangrui Peng Shanyu Xiao Gang Yao
Beijing National Research Center for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing, China
{yrj17, xiaosy19, yg19}@mails.tsinghua.edu.cn, penglr@tsinghua.edu.cn

In order to investigate more configurations of the proposed primitive representation learning framework, we conduct supplementary experiments including different input sources, different fusion strategies for PREN (Primitive REpresentation learning Network), and different types of encoders in PREN2D (PREN with the 2D attention mechanism). More visualization results and error analysis are also provided.

## 1. Different input sources for PREN

EfficientNet-B3 [1] is used as the feature extraction module. Let $F_i$ denote the output feature maps of the $i$-th convolutional block, we compare PRENs that learn primitive representations from feature maps with various resolutions as inputs.

As shown in Table 1, the feature maps learned by the final convolutional block $F_7$ plays the most important role in recognition performance. Nevertheless, feature maps $F_3$ and $F_5$ with higher resolutions can help the model learn more local spatial information. The model that uses all three features achieves the highest accuracy on three test sets.

Table 1. Word accuracy (%) of PRENs that learn primitive representations from feature maps with various resolutions as inputs. $F_i$ denotes the output of the $i$-th convolutional block.

| Features | IIIT5k | IC03 | IC13 | SVTP | CUTE |
|---|---|---|---|---|---|
| $F_3$ | 77.6 | 84.8 | 85.1 | 66.5 | 53.8 |
| $F_5$ | 90.3 | 93.8 | 93.3 | 80.0 | 76.0 |
| $F_7$ | 91.1 | 93.7 | 94.5 | **82.5** | 79.2 |
| $F_3 + F_5$ | 90.1 | 93.3 | 92.9 | 78.5 | 72.9 |
| $F_3 + F_7$ | 90.6 | **94.5** | 94.4 | 81.2 | 79.5 |
| $F_5 + F_7$ | 90.9 | 93.9 | 94.3 | 80.9 | 78.5 |
| $F_3 + F_5 + F_7$ | **91.8** | 93.9 | **94.7** | 81.7 | **81.3** |

Table 2. Word accuracy (%) of PRENs with various fusion strategies for visual text representations.

| Fusion | IIIT5k | IC03 | IC13 | SVTP | CUTE |
|---|---|---|---|---|---|
| Summation | **91.8** | 93.9 | **94.7** | **81.7** | **81.3** |
| Concat | 91.1 | 93.7 | 94.0 | **81.7** | 78.8 |
| Gated Unit | 91.6 | **94.1** | 93.7 | 81.1 | 80.2 |

## 2. Comparison of various fusion strategies for visual text representations

For PREN, the fused visual text representations $Y$ is obtained by fusing the two types of visual text representations $Y_1$ and $Y_2$, which are generated from primitive representations learned by pooling aggregators and weighted aggregators, respectively. We study three fusion strategies, i.e., summation, concatenation and gated unit. As shown in Table 2, the model with the summation fusion strategy can achieve the best recognition accuracy on most test sets.

## 3. Comparison of the encoder

For the encoder of PREN2D, we propose a modified self-attention mechanism by using $3 \times 3$ convolutions before the computation of attention scores. We compare our method with the model that uses an original self-attention network as an encoder. As shown in Table 3, the proposed modified self-attention network can outperform the original self-attention network on most test sets.

Table 3. Word accuracy (%) of PREN2Ds with different encoders.

| Encoder | IIIT5k | IC03 | IC13 | SVTP | CUTE |
|---|---|---|---|---|---|
| self-attention | 95.1 | **95.3** | 95.9 | 86.2 | 88.5 |
| Ours | **95.6** | **95.3** | **96.0** | **86.7** | **88.9** |

## 4. More visualization results

Fig. 1 visualizes attention scores output by Baseline2D and PREN2D with respect to another three samples. Base-

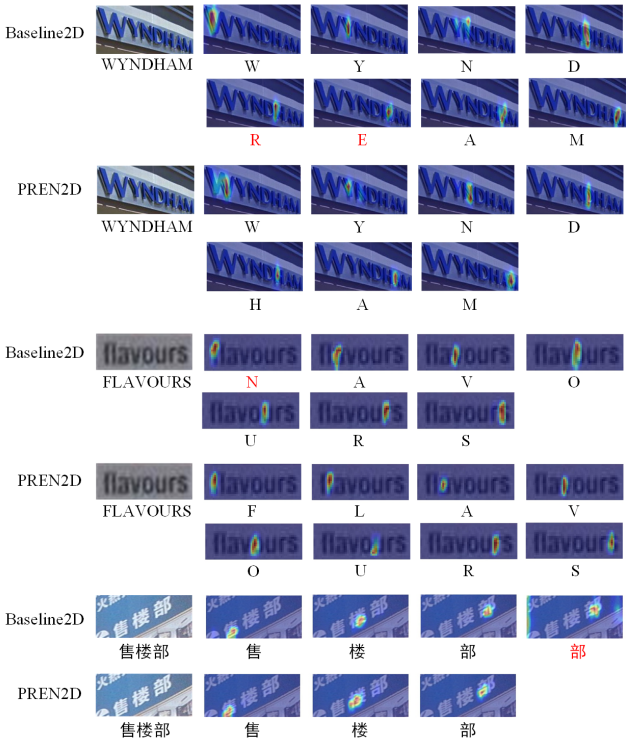


Figure 1. Visualization of attention scores generated by Baseline2D and PREN2D for two English text images and a Chinese text image. Wrongly recognized characters are marked in red.

line2D recognizes one more or one less character, which indicates that the misalignment problem may occur. In contrast, PREN2D can generate more accurate attention areas and make correct predictions.

In experiments of the submitted paper, we have found that too many primitive representations cannot have a positive effect on recognition performance. We draw heatmaps generated by a weighted aggregator that learns 9 primitive representations. As shown in Fig. 2, some heatmaps are nearly identical, thus cannot provide additional information that is helpful for recognition.

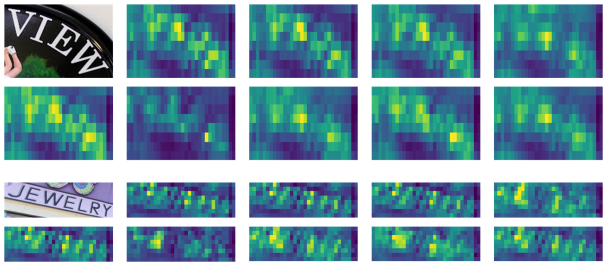


Figure 2. Heatmaps generated by the weighted aggregator that learns $n = 9$ primitive representations.

## 5. Error Analysis

Fig. 3 shows some failure cases of PREN and PREN2D. Similar characters may be confused by PREN. For PREN2D, out-of-vocabulary words are sometimes recognized as other words, e.g., PREN2D recognizes "WESC" as "WEST". Complex fonts and low-quality images are still challenging for both PREN and PREN2D.


Figure 3. Examples of failure cases. Sub-figures (a), (b) and (c) show the examples wrongly recognized by PREN, PREN2D, both PREN and PREN2D, respectively.

## References

[1] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019. 1