STAR-Net: A SpaTial Attention Residue Network for Scene Text Recognition

Wei Liu[1]
wliu@cs.hku.hk
Chaofeng Chen[1]
cfchen@cs.hku.hk
Kwan-Yee K. Wong[1]
kykwong@cs.hku.hk
Zhizhong Su[2]
suzhizhong@baidu.com
Junyu Han[2]
hanjunyu@baidu.com

[1] Department of Computer Science
The University of Hong Kong, HK

[2] Institution of Deep Learning
Baidu Inc, Beijing

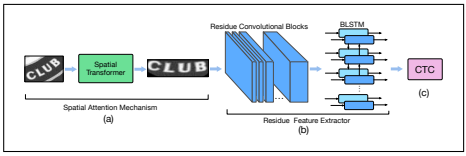

Figure 1: Overview of our STAR-Net for scene text recognition. (a) Spatial attention mechanism. (b) Residue feature extractor. (c) Connectionist Temporal Classification.

In this paper, we present a novel SpaTial Attention Residue Network (**STAR-Net**) for recognising scene texts. The overall architecture of our STAR-Net is illustrated in fig. 1. Our STAR-Net emphasises the importance of representative image-based feature extraction from text regions by the spatial attention mechanism and the residue learning strategy. It is by far the deepest neural network proposed for scene text recognition.

Spatial Attention Mechanism The spatial transformer [2] is responsible for introducing the spatial attention mechanism (see fig. 2(a)). A localisation network is used to determine the transformation parameters $\theta(I)$ of the original text image. Based on these parameters, a sampler locates sampling points on the input image which explicitly define the text region to be unwarped. Finally, an interpolator generates the output image by interpolating the intensity values of the four pixels nearest to each sampling point.

Residue Learning Strategy To fully exploit the potential of convolutional layers and build up a powerful deep feature encoder, we employ residue convolutional blocks [1] (see fig. 2(b)) along with Long Short-Term Memory to extract informative features from the rectified text regions.

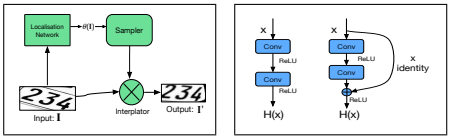

Figure 2: Structures of the spatial transformer, plain and residue convolutional blocks.

Experimental Results Experiments conducted on the SVT-Perspective dataset show that our STAR-Net outperforms other state-of-the-art methods on both lexicon-based and lexicon-free recognition. Besides STAR-Net, we also evaluate three other network architectures (CRNN, STA-CRNN and R-Net) to demonstrate the effectiveness of each component in our STAR-Net. More details are in the experiment part of the paper.

Wang *et al.*	40.5	26.1	-
Mishra *et al.*	45.7	24.7	-
Wang *et al.*	40.2	32.4	-
Phan *et al.*	75.6	67.0	-
Shi *et al.*	91.2	77.4	71.8
CRNN	92.6	72.6	66.8
STA-CRNN	93.0	80.5	69.3
R-Net	93.0	83.6	70.9
STAR-Net	**94.3**	**83.6**	**73.5**

Table 1: Scene text recognition accuracies (%) on SVT-Perspective dataset.

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[2] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015.