Show，Attend and Read:
A Simple and Strong Baseline for Irregular Text Recognition

Hui Li∗
School of Computer Science, The University of Adelaide, Australia, 5000

Peng Wang∗
School of Computer Science, Northwestern Polytechnical University, China

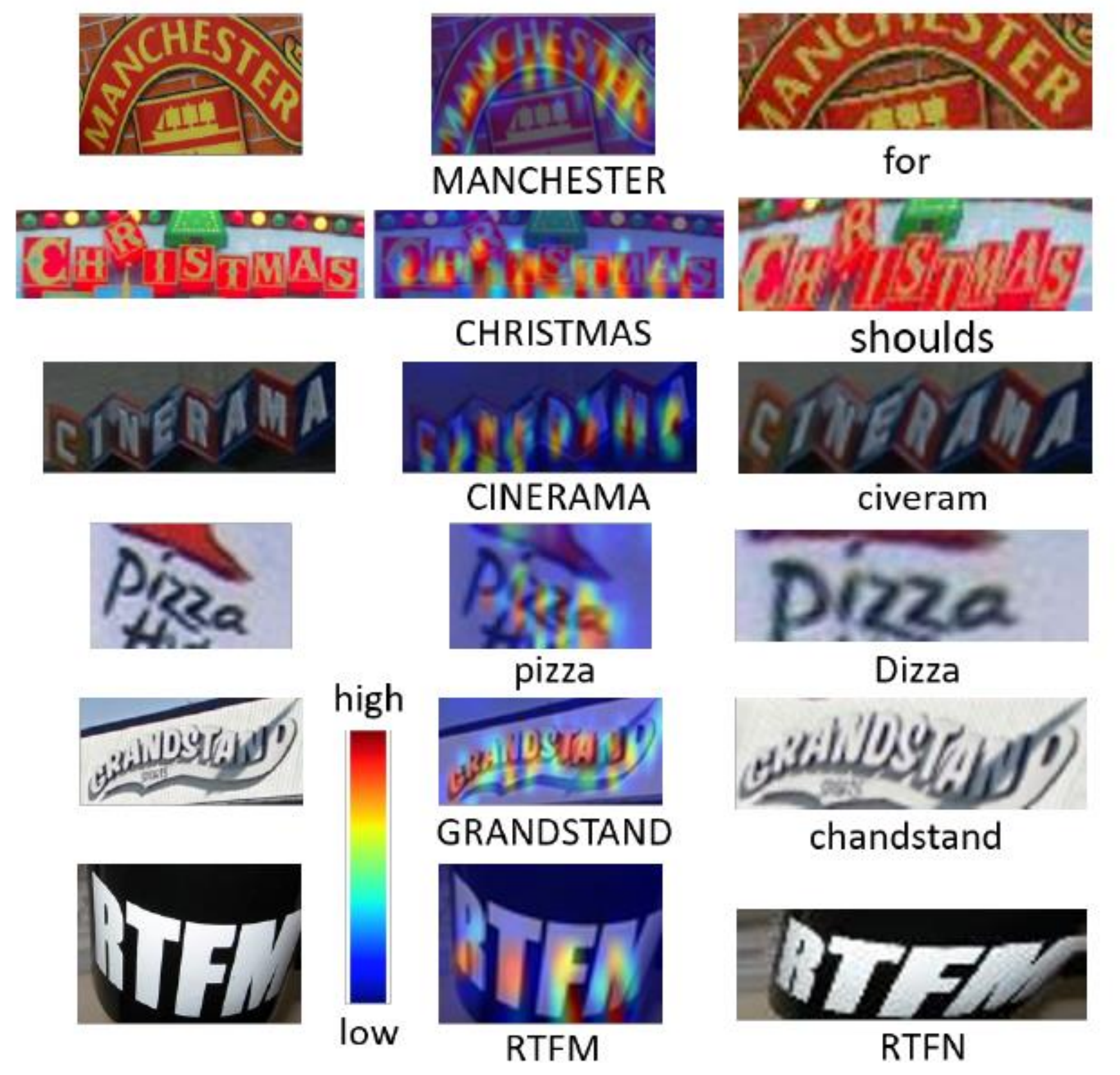
Chunhua Shen
School of Computer Science, The University of Adelaide, Australia, 5000

Guyu Zhang
School of Computer Science, Northwestern Polytechnical University, China

Introduction

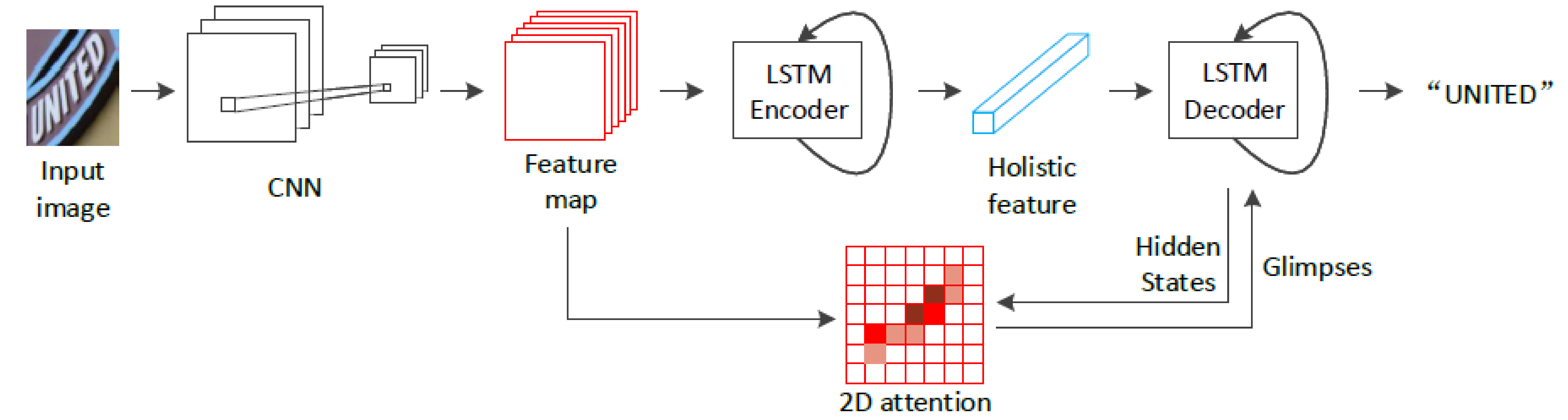
In this work, we propose a simple and strong framework for irregular text recognition. The main contributions of this work is three-fold:

1. We setup an easy-to-implement strong baseline for recognizing irregular text in natural scene images, which is made up of off-the-shelf neural components such as CNNs, LSTMs and attention mechanisms. The proposed model can be trained end-to-end without pre-training. All the training examples are synthetic or from public real data. We will release the code and data used for training.
2. Compared to existing irregular text recognizers, our proposed approach does not rely on sophisticated designs (including spatial transformation, hierarchical attention or multi-directional encoding) to handle text distortions. Alternatively, we simply use a 2D attention mechanism to deal with irregular text, which selects local features for individual characters. Moreover, our proposed attention module does not require additional pixel-level or character-level supervision information, which is weakly supervised by the cross-entropy loss on the final predictions. The attention mechanism is also tailored to consider neighborhood information and boosts the recognition performance.
3. Note that many irregular text recognizers perform relatively worse on regular text. In contrast, due to its flexibility and robustness, the proposed approach not only significantly outperforms existing approaches on irregular text, but also achieves favorable performance on regular text.

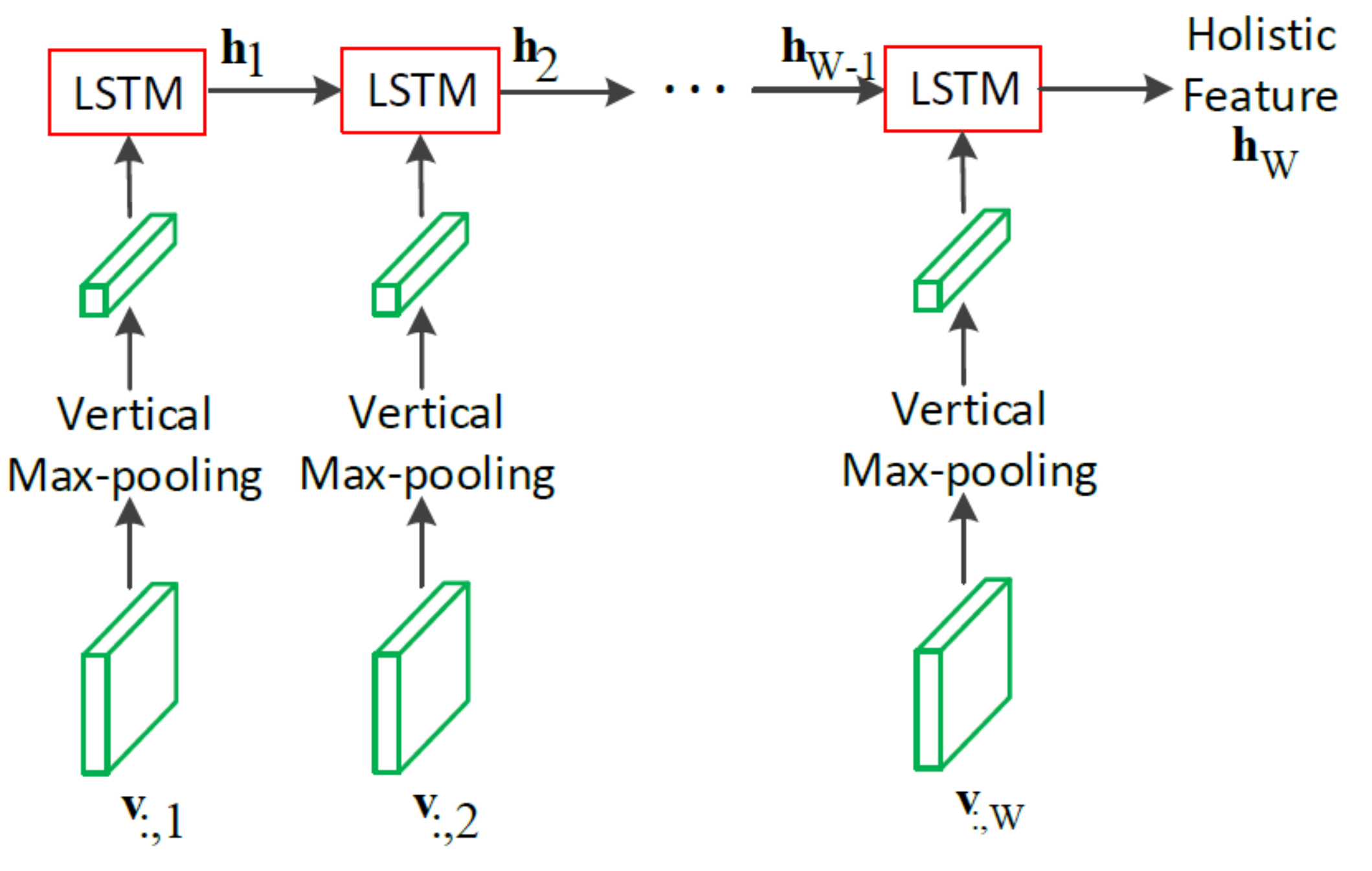


(a) Original Image (b) 2D attention (ours) (c) Rectification (Shi et al. 2018)

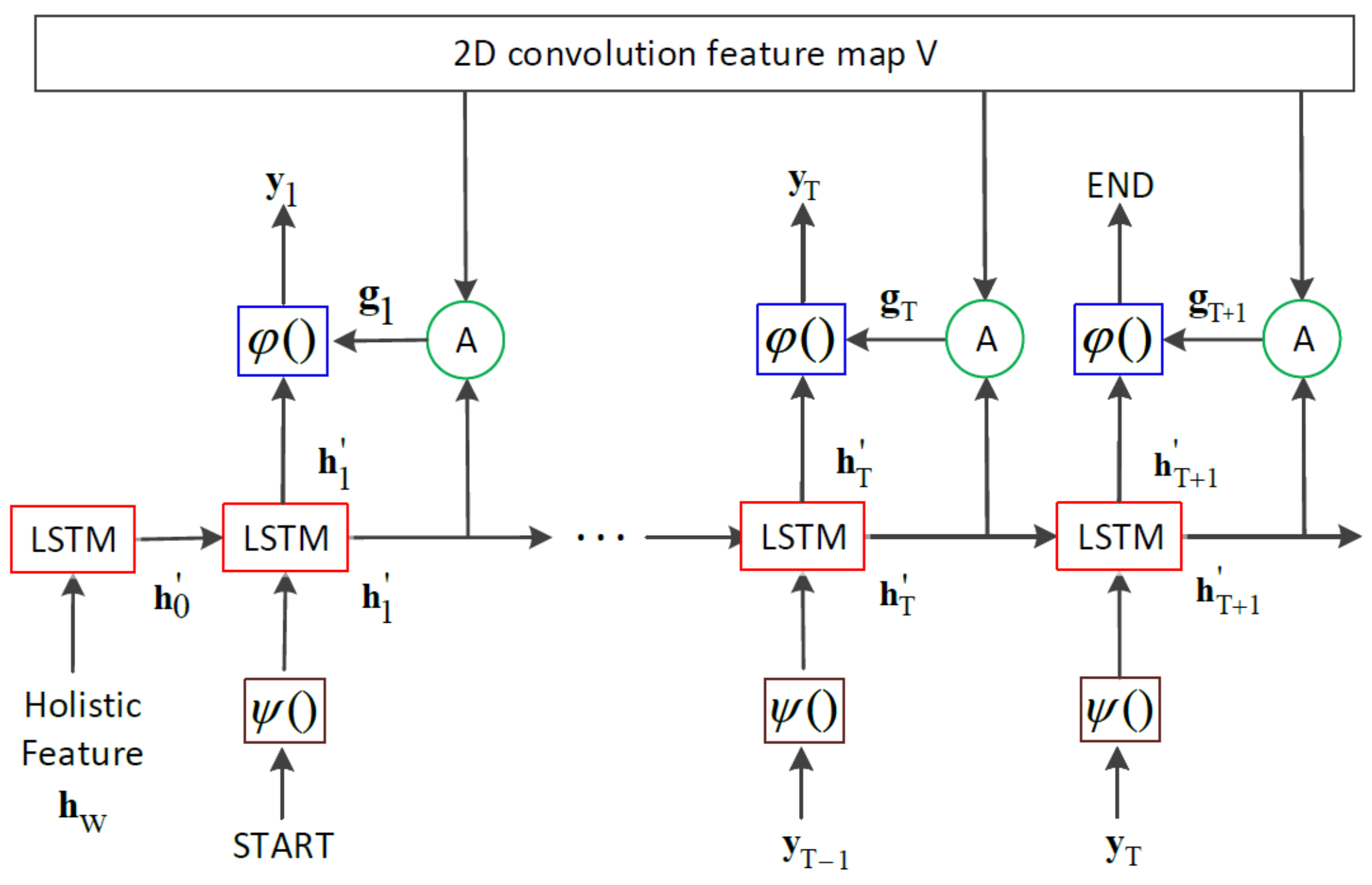
Framework



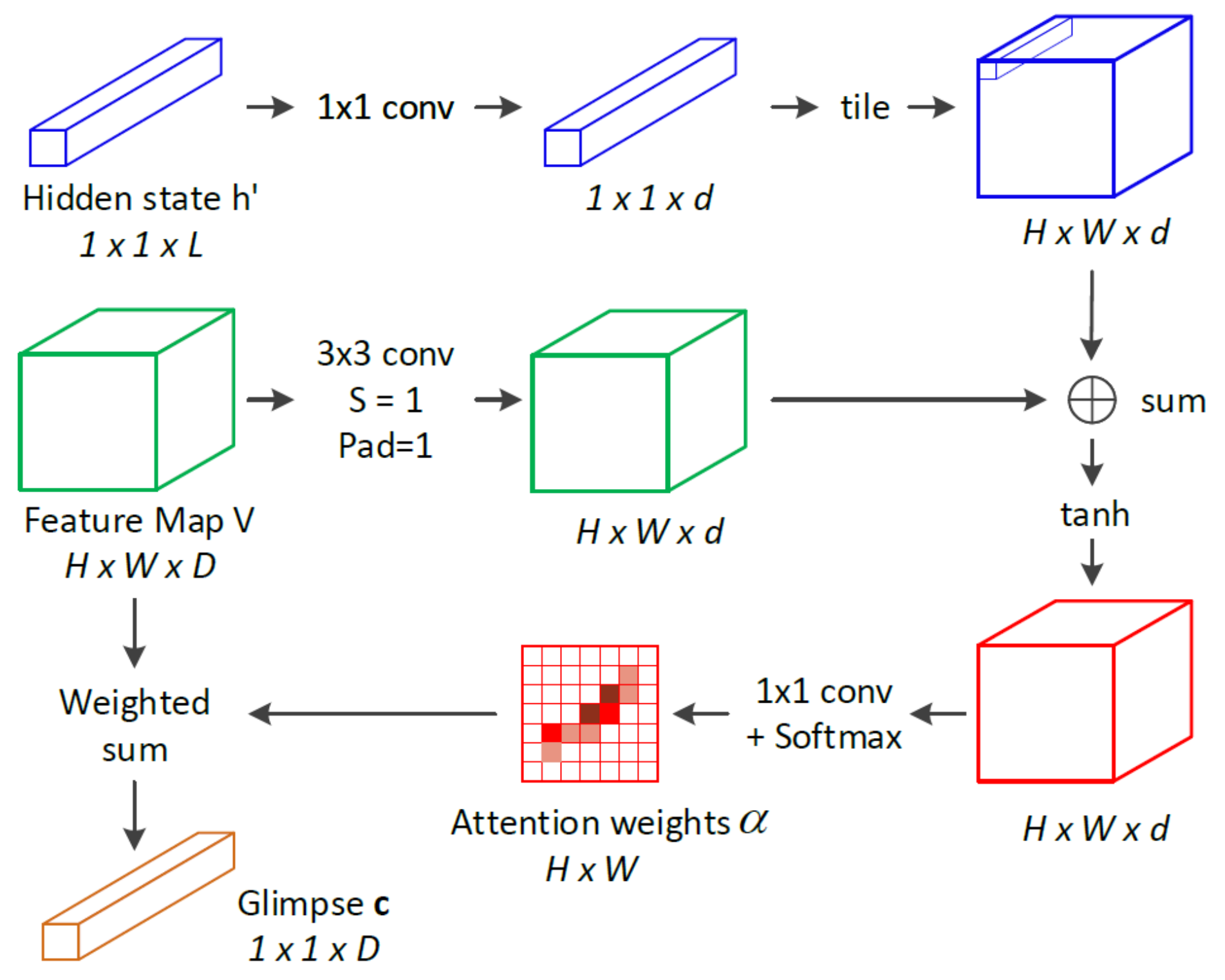
Encoder



Decoder



2D Attention Module



Results

Method	Regular Text						Irregular Text					
	IIIT5K			SVT		IC13	IC15	SVTP			CT80	COCO-T
	50	1k	None	50	None	None	None	50	Full	None	None	None
(Wang, Babenko, and Belongie 2011)	–	–	–	57.0	–	–	–	40.5	21.6	–	–	–
(Mishra, Alahari, and Jawahar 2012b)	64.1	57.5	–	73.2	–	–	–	45.7	24.7	–	–	–
(Phan et al. 2013)	–	–	–	73.7	–	–	–	75.6	67.0	–	–	
(Yao et al. 2014)	80.2	69.3	–	75.9	–	–	–	–	–	–	–	
(Jaderberg et al. 2015a)	97.1	92.7	–	95.4	80.7	90.8	–	–	–	–	42.7	
(He et al. 2016b)	94.0	91.5	–	93.5	–	–	–	–	–	–	–	
(Lee and Osindero 2016)	96.8	94.4	78.4	96.3	80.7	90.0	–	–	–	–	–	
(Wang and Hu 2017)	98.0	95.6	80.8	96.3	81.5	–	–	–	–	–	–	
(Shi et al. 2016)	96.2	93.8	81.9	95.5	81.9	88.6	–	91.2	77.4	71.8	59.2	–
(Liu et al. 2016)	97.7	94.5	83.3	95.5	83.6	89.1	–	*94.3*	83.6	73.5	–	
(Shi, Bai, and Yao 2017)	97.8	95.0	81.2	97.5	82.7	89.6	–	92.6	72.6	66.8	54.9	–
(Yang et al. 2017)*	97.8	96.1	–	95.2	–	–	–	93.0	80.2	75.8	69.3	–
(Cheng et al. 2017)*	99.3	97.5	87.4	97.1	85.9	93.3	70.6	92.6	81.6	71.5	63.9	–
(Liu et al. 2018)*	97.0	94.1	87.0	95.2	–	92.9	–	–	–	–	–	
(Liu, Chen, and Wong 2018)*	–	–	92.0	–	85.5	91.1	74.2	–	–	*78.9*	–	*59.3*
(Bai et al. 2018)*	*99.5*	97.9	88.3	96.6	87.5	**94.4**	73.9	–	–	–	–	
(Cheng et al. 2018)	**99.6**	98.1	87.0	96.0	82.8	–	68.2	94.0	*83.7*	73.0	76.8	
(Shi et al. 2018)	**99.6**	**98.8**	*93.4*	**99.2**	**93.6**	91.8	*76.1*	–	–	78.5	*79.5*	
SAR (Ours)	99.4	*98.2*	**95.0**	*98.5*	*91.2*	*94.0*	**78.8**	**95.8**	**91.2**	**86.4**	**89.6**	**66.8**

Attention Weights Visualization