Mutually Guided Dual-Task Network for Scene Text Detection

Mengbiao Zhao[1,2] , Wei Feng[1,2], Fei Yin[1,2] , Xu-Yao Zhang[1,2], Cheng-Lin Liu[1,2,3]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

[2]University of Chinese Academy of Sciences

[3]CAS Center for Excellence of Brain Science and Intelligence Technology

中国科学院 CHINESE ACADEMY OF SCIENCES

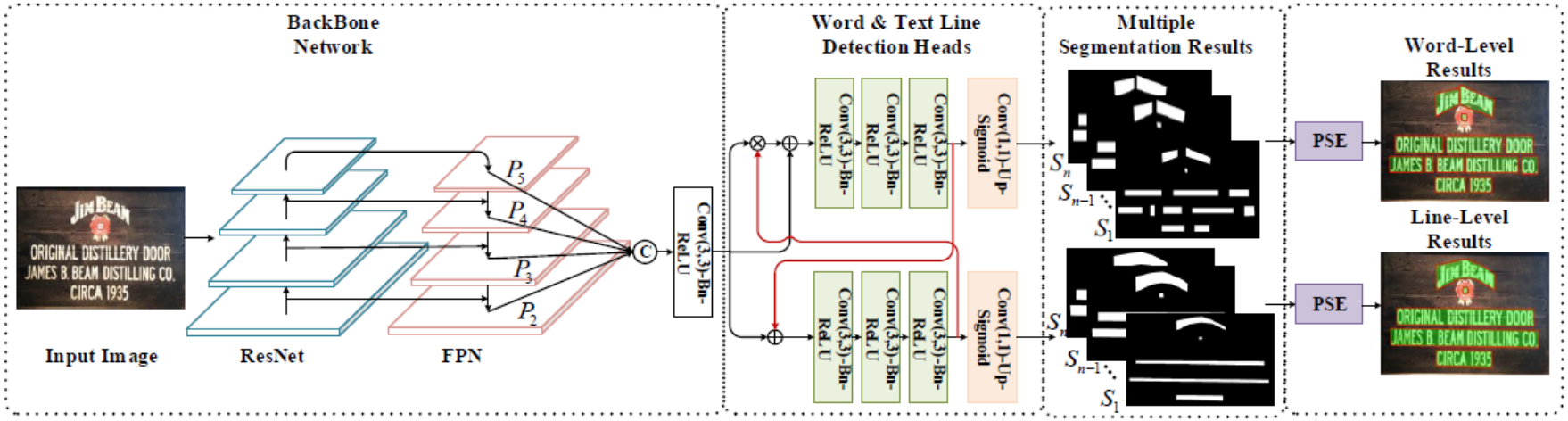
中国科学院自动化研究所 Institute of Automation, Chinese Academy of Sciences

Outline

Introduction

- There are two mainstream annotation formats for scene text datasets: word-level and line-level annotations.
- In previous works, word detection and text-line detection are usually treated separately.
- Word-level and line-level detection are closely related.

System Overview

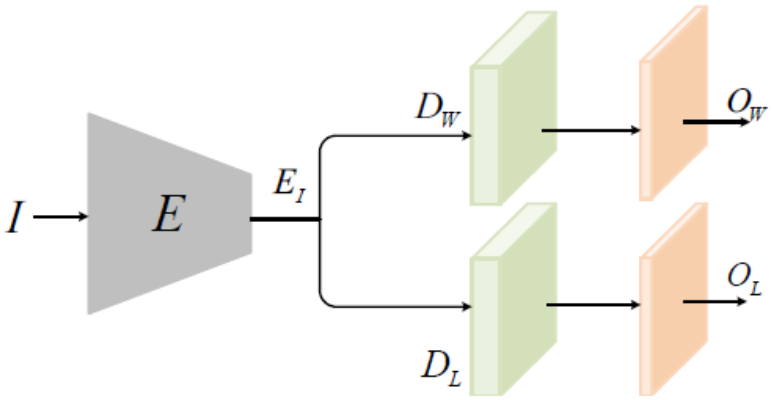


➢Architecture

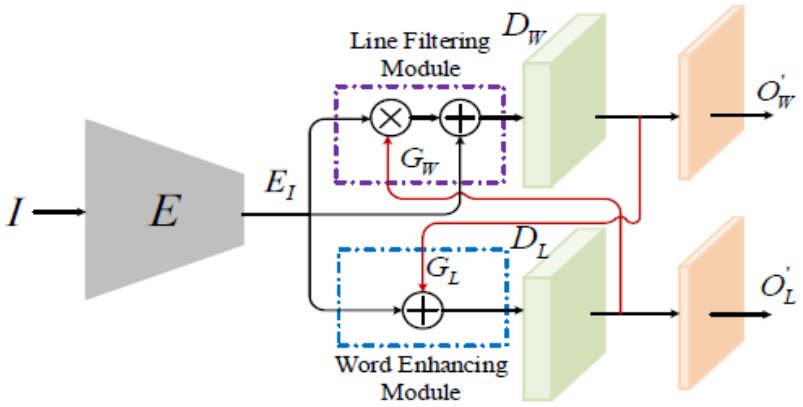
- A backbone network (ResNet-50 FPN) for feature extraction.
- Two detection heads for words and text-lines detection, respectively.
- Two novel modules for the mutual guidance of the two tasks.

Mutual Guidance Strategy

For ease of analysis, we divide the training process into two stages.



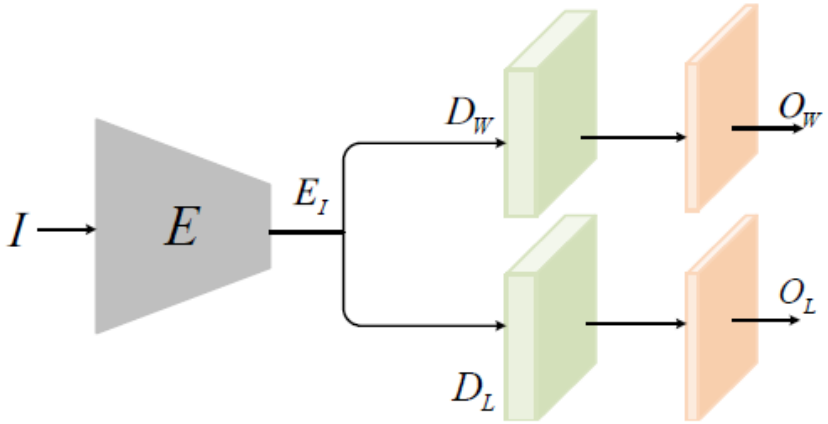
Stage 1



Stage 2

Mutual Guidance

- Stage 1



E : backbone network

E_I : original features

D_W : word detector

D_L : text-line detector

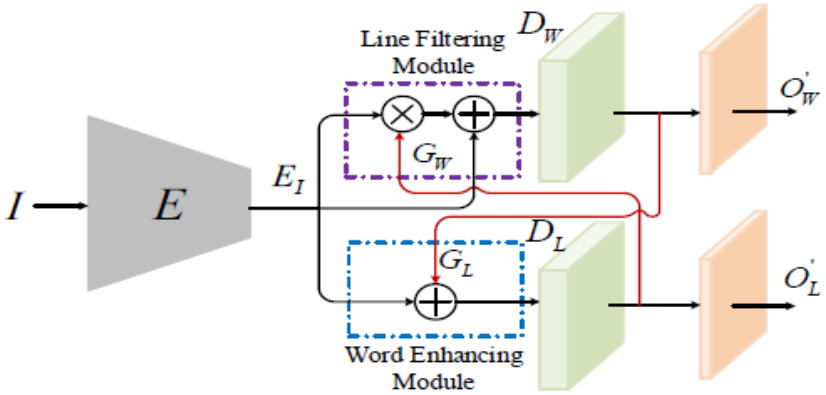
O_W : output results of word detector

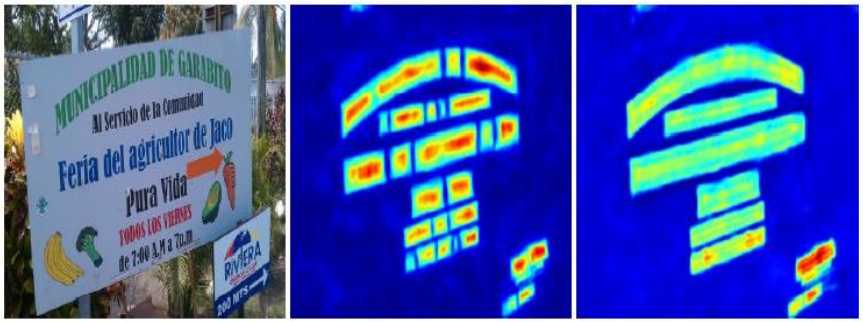
O_L : output results of text-line detector

$$\begin{cases} O_W = D_W(E_I), \\ O_L = D_L(E_I). \end{cases}$$

Mutual Guidance

➢ Stage 2





- Line filtering modules.

$$E_I^{'} = E_I \odot G_W + E_I,$$

- Word enhancing modules.

$$E_I^{'} = E_I + G_L,$$

$$\begin{cases} O_W^{'} = D_W(E_I, G_W), \\ O_L^{'} = D_L(E_I, G_L). \end{cases}$$

Loss Function

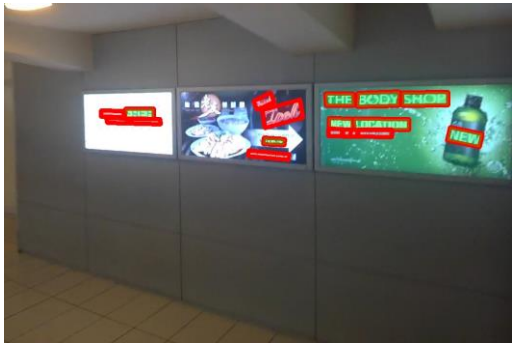
- We use pair of datasets, one with word-level ground truth Y_W and one with line-level ground truth Y_L .
- For a data batch with Y_W, we just compute the dice coefficient loss between Y_W and its two stages' outputs O_W and O'_W.

$$\mathcal{L} = \sum_{t \in \{W,L\}} \sum_{X \in \{O_t, O'_t\}} b_t \cdot \mathcal{L}_{dice}(X, Y_t),$$

- Where b_t represents the category of the current data batch.
- If a data batch has ground truth Y_W only, then $b_W = 1, b_L = 0$ and vice verse.

Experiments

- Datasets
  - ICDAR2015: Word-level annotated dataset.
  - CTW1500: Line-level annotated dataset.



Experiments

➢ Models

- ***Baseline:*** The basic detector trained with word-level and line-level annotated data separately
- ***Baseline + joint***: The basic detector jointly trained with word-level and line-level annotated data.
- ***Dual-task***: Our proposed dual-task network jointly trained with word-level and line-level annotated data.
- ***Dual-task + guidance***: Our proposed dual-task network jointly trained with word-level and line-level annotated data, and the mutual guidance strategy added.

Experiments

- Ablation studies on ICDAR2015

Method	P	R	F
Baseline [3]	81.5	79.7	80.6
Baseline+joint [3]	82.32	76.45	79.28
Dual-task	**87.41**	74.91	80.68
Dual-task+guidance	82.08	**80.98**	**81.53**

- The basic detector jointly trained with two datasets yields deteriorated performance.
- The dual-task network leads to an improved performance.
- The dual-task network trained with mutual guidance yields the best detection performance.

Experiments

- Comparison with State-of-the-Art Methods.
  - Detection results on ICDAR2015.

Method	Ext	P	R	F
EAST [27]	-	83.57	73.47	78.2
PixelLink [8]	-	82.9	81.7	82.3
TextBoxes++ [28]	✓	87.2	76.7	81.7
DDR [1]	-	82.0	80.0	81.0
FOTS [29]	✓	88.84	82.04	85.31
Mask TextSpotter [30]	✓	**91.6**	81.0	86.0
TextField [31]	✓	84.3	80.5	82.4
TextSnake [32]	✓	84.9	80.4	82.6
PSENet [3]	-	81.5	79.7	80.6
PSENet [3]	✓	86.9	84.5	85.7
Our Method	-	82.08	80.98	81.53
Our Method	✓	88.60	**84.54**	**86.52**

Experiments

- Comparison with State-of-the-Art Methods.
  - Detection results on CTW1500.

Method	Ext	P	R	F
CTPN [12]	-	60.4	53.8	56.9
SegLink [33]	-	42.3	40.0	40.8
CTD+TLOC [4]	-	77.4	69.8	73.4
TextSnake [32]	✓	67.9	**85.3**	75.6
Wang et al. [3]	-	80.1	80.2	80.1
TextField [31]	✓	83.0	79.8	81.4
PSENet [3]	-	80.57	75.55	78.0
PSENet [3]	✓	84.84	79.73	82.2
Our Method	-	81.48	78.42	79.92
Our Method	✓	**85.59**	80.21	**82.81**

Experiments

- Some examples of text detection.

ICDAR2015 CTW1500

- Each image can get two formats of detection results from two detection heads.

Conclusions

- **Propose a text detection method that can perform both word-level and line-level text detection.**
  - Dual-task network.
- **Propose two novel modules for the mutual guidance of the two tasks.**
  - Line filtering module.
  - Word enhancing module.
- **Proposed method has achieved competitive performance.**
- **Future works**
  - Weakly-supervised training.
  - Adding character-level detection.

Thanks & Question