IEEE GlobalSIP

Efficient Segmentation-Aided Text Detection For Intelligent Robots

Junting Zhang, Yuewei Na, Siyang Li, C.-C. Jay Kuo
University of Southern California

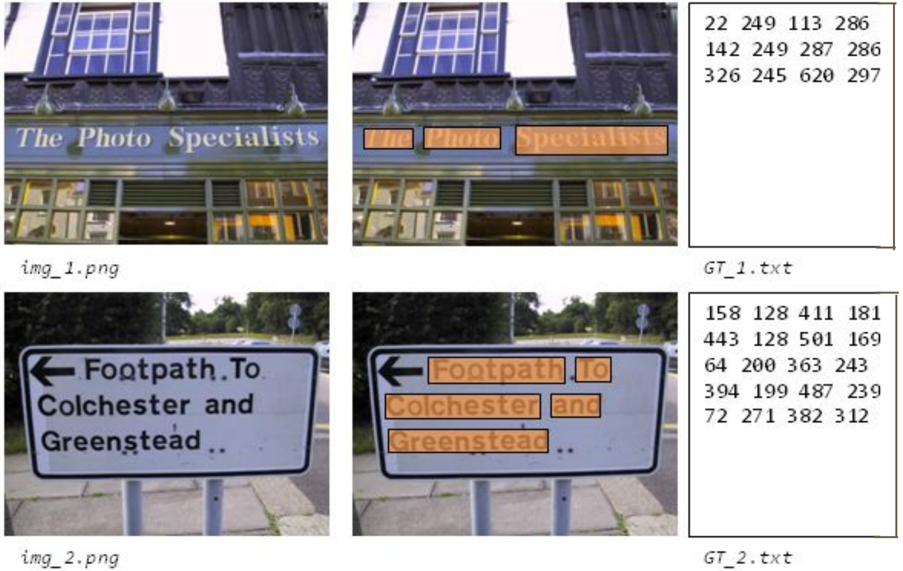
USC University of Southern California

Outline

- Problem Definition and Motivation
- Related Work
  - Detection-based approaches
  - Casting the detection problem as a semantic segmentation problem
- Segmentation-aided Text Detection
  - Methodologies
  - Experimental Results
- Conclusions and Future Work

Problem Definition

Text Detection: Given an image, **word-level** bounding boxes should be produced.



```
22 249 113 286
142 249 287 286
326 245 620 297
```

```
158 128 411 181
443 128 501 169
64 200 363 243
394 199 487 239
72 271 382 312
```

Original Image

Visualization of G.T.

G.T. Text file

Motivation

- ADAS and Robot Vision

Motivation

- Visual Translation:
  - Current app requires user to align the text manually

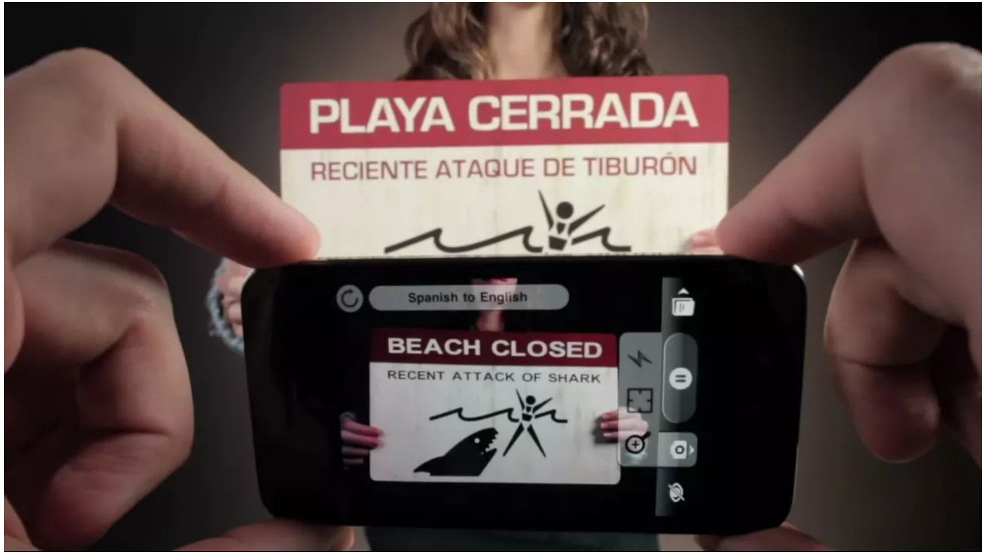
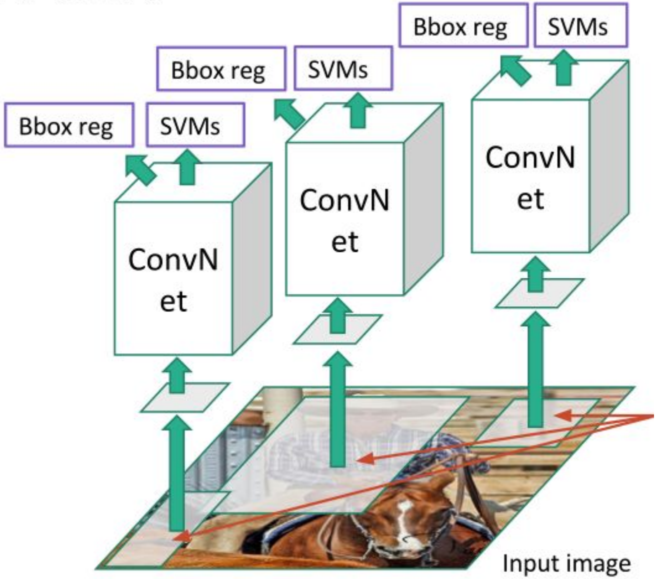


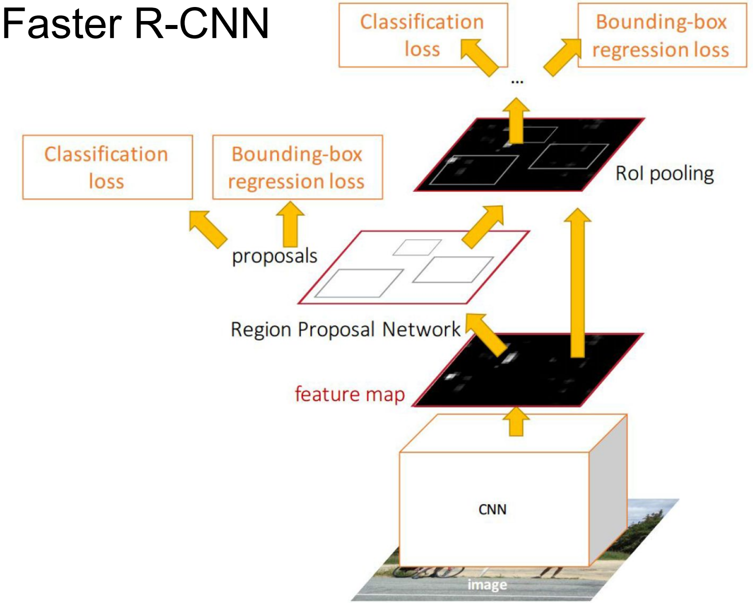
Image credit: Google Translate

Extending the generic object detectors

R-CNN



Faster R-CNN



Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.

Extending the object detectors

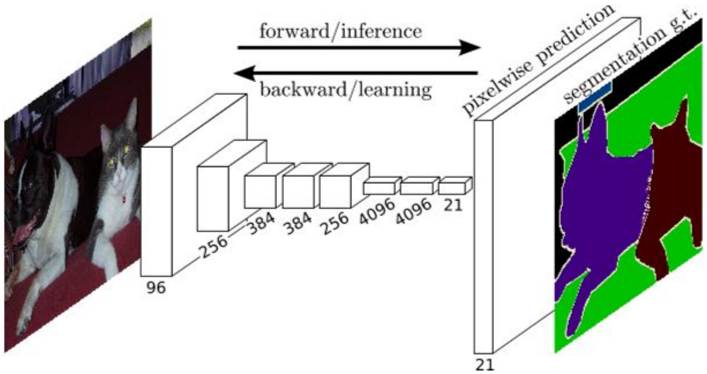
Predict text boxes directly by sliding a window over the convolutional features.

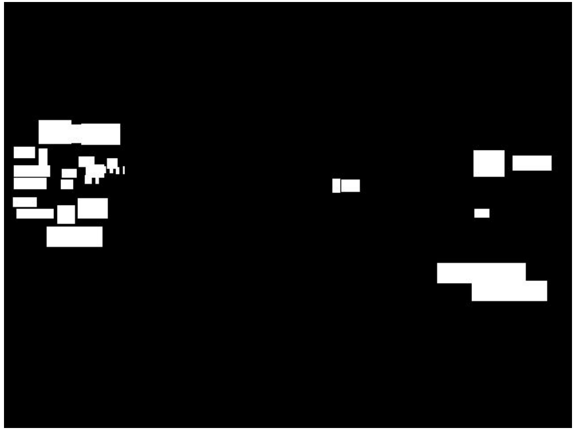
Predictions are made based on **local regions** in the image without enough contextual cues.

When the environment becomes more challenging, they are not robust against text-like patterns such as fences, brick wall, windows, leaves, etc..

Tedious Post-processing is usually needed to **remove false positives**.

Casting the detection problem as a semantic segmentation problem

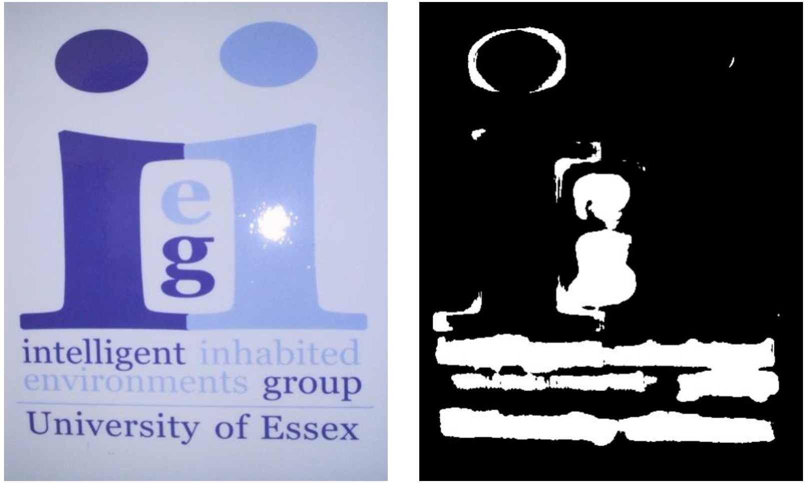




Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

Segmentation Networks

- FCNs usually have **large receptive fields**
  - Advantage: More context information is considered, and the predictions are robust to text-like patterns
  - Disadvantage: Hard to do the fine-scale detection, i.e. unable to separate the lines and words
    - Recall: The desired output should be **word-level** bounding boxes



Problems of Segmentation-based Approaches:

- To produce fine-scale detection results, one has to use:
  - Cascaded system to process each ROI separately -->Not efficient!
  - Extra annotations, eg. word center, character center. -->Too costly!

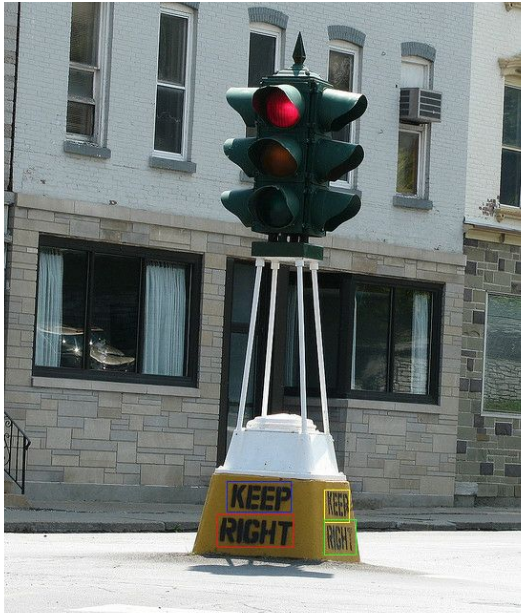
WHY NOT Combine DET and SEG?

Essential Intuition

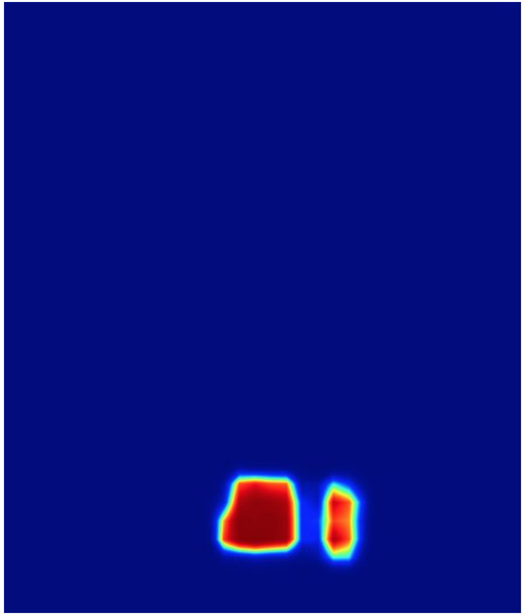
- Detection network is error prone to text-like patterns, but good at predicting accurate bounding boxes for each individual word
    - Detection net is wasting efforts at wrong regions!
- Segmentation network is robust to clutter backgrounds, but unable to separate individual words
    - Segmentation net is suitable for ROI finding.
- Let's use segmenter output to **guide** the detector, so that it can pay **attention** to the correct regions.

Text Attention Map (TAM)

- TAM is a heat map that indicating the probability of existence of text.
- A TAM can be obtained by training a FCN with text region mask.



Input Image with ground truth bounding boxes.



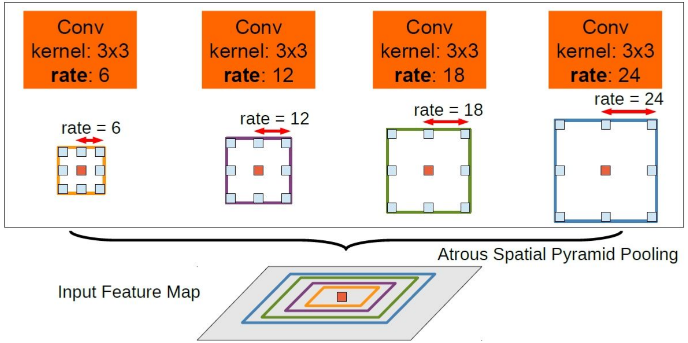
TAM.
Red means higher confidence score.

How to obtain a good TAM for an input image?

Atrous Spatial Pyramid Pooling (ASPP) Layer

ASPP produces **multi-scale representations** by combining feature responses from parallel atrous convolution layers with **different sampling rates**.

Dilation always brings more global view!

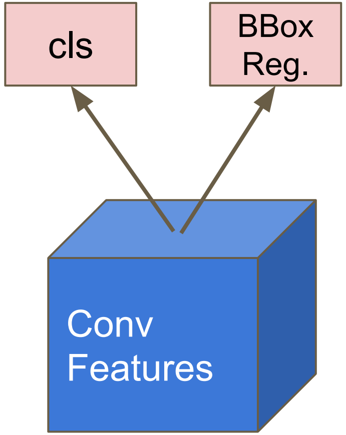


Atrous Spatial Pyramid Pooling Layer

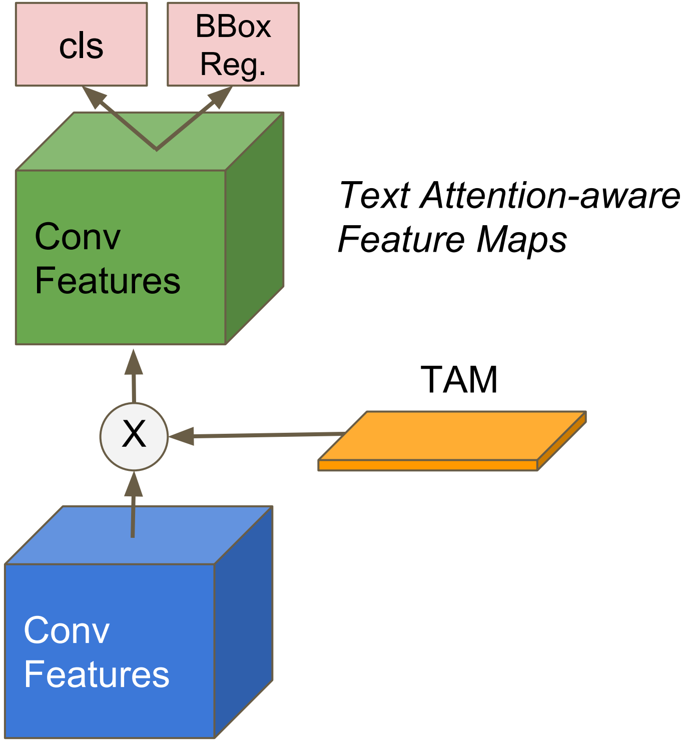
Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." *arXiv preprint arXiv:1606.00915* (2016).

How to use the TAM to guide the detector?

Multiplicative Gating

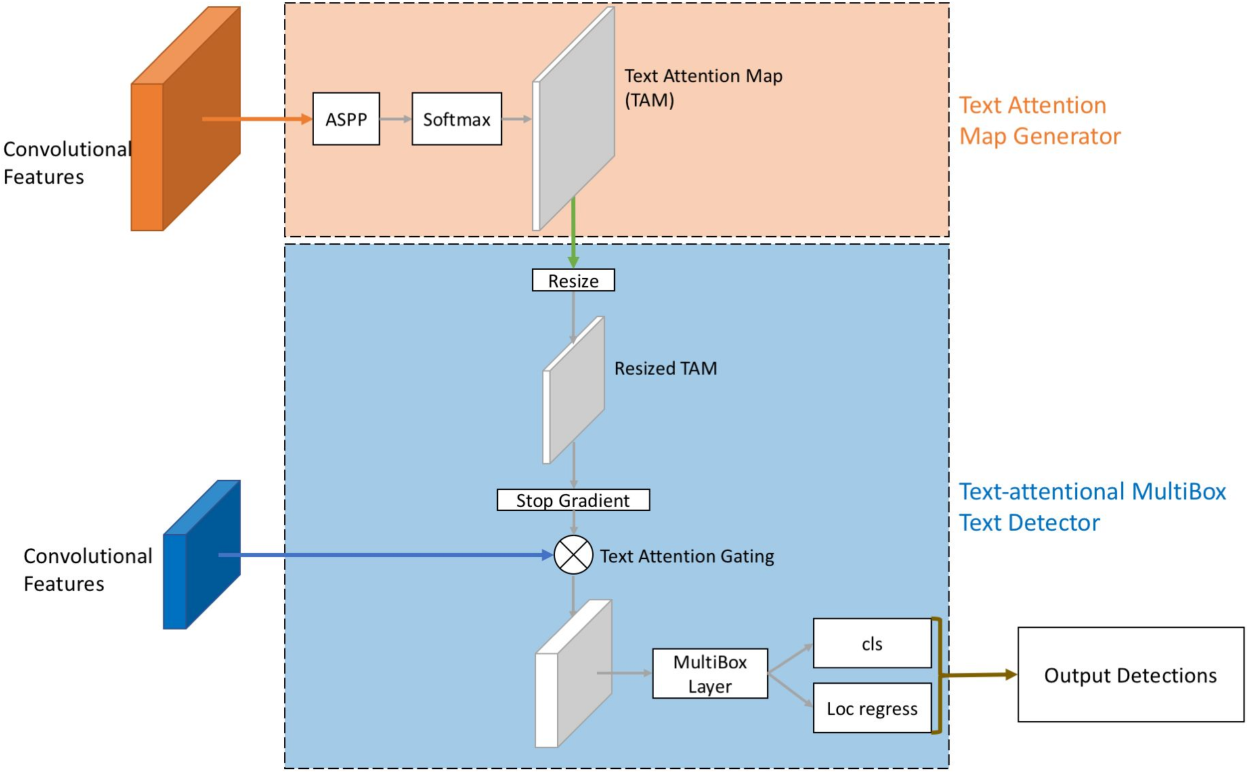


Detector without Text Attention Gating



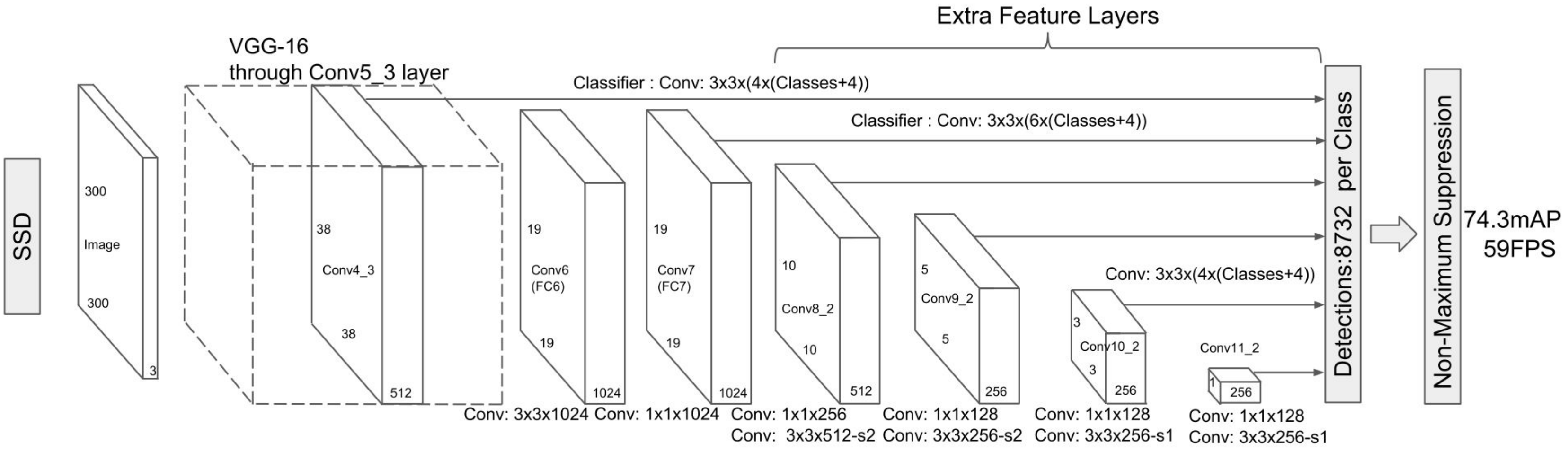
Detector with Text Attention Gating

Text Attention Gating



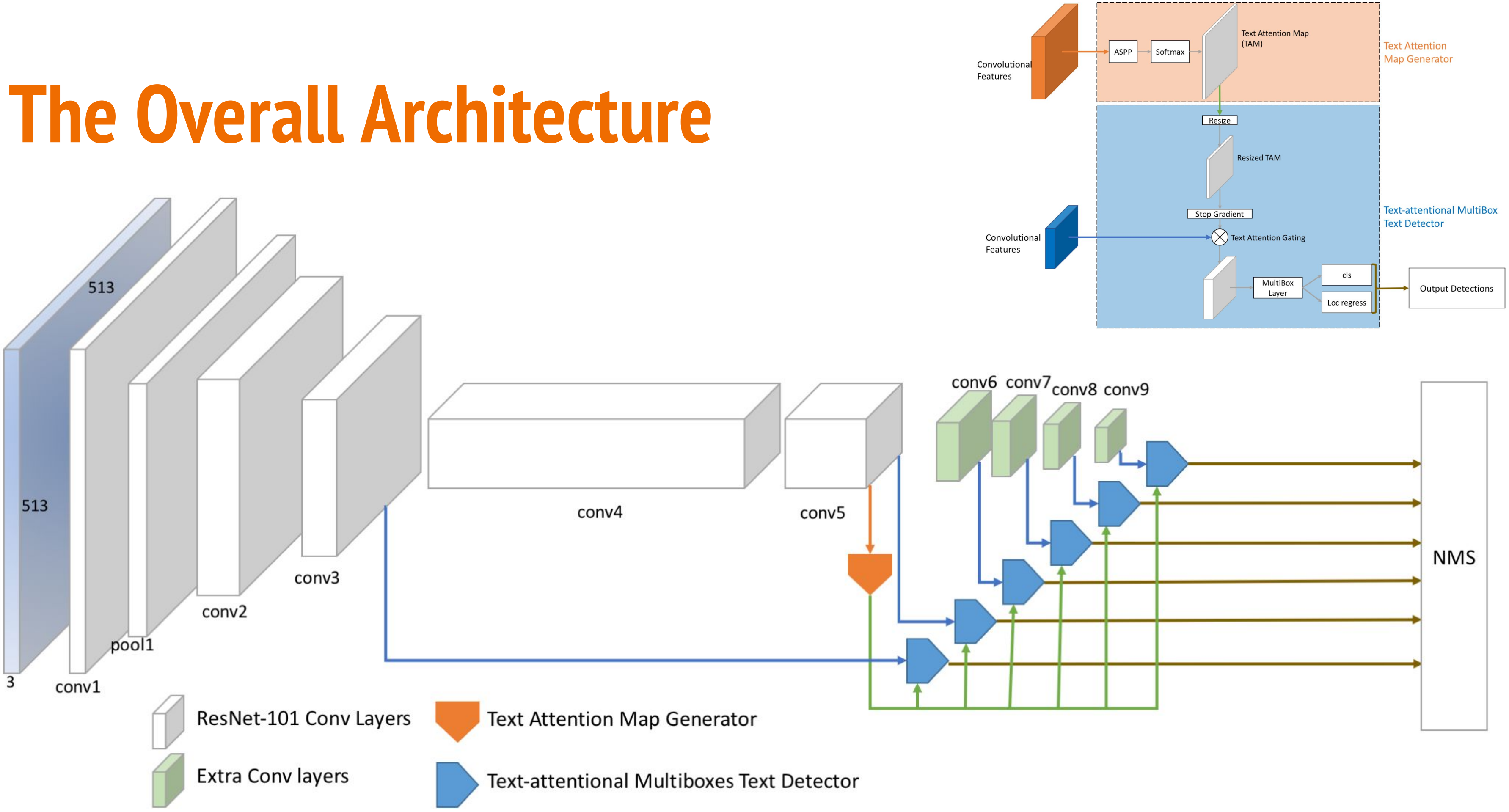
Which Detector to use?

Single Shot MultiBox Detector (SSD): The most efficient detector up to date.



Liu, Wei, et al. "Ssd: Single shot multibox detector." *European conference on computer vision*. Springer, Cham, 2016.

The Overall Architecture



Experimental Results

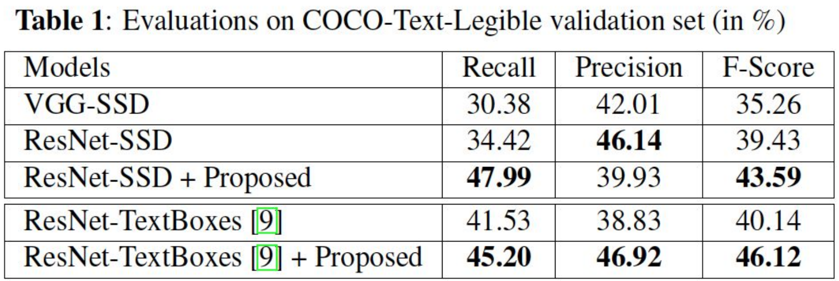


Table 1: Evaluations on COCO-Text-Legible validation set (in %)

Models	Recall	Precision	F-Score
VGG-SSD	30.38	42.01	35.26
ResNet-SSD	34.42	**46.14**	39.43
ResNet-SSD + Proposed	**47.99**	39.93	**43.59**
ResNet-TextBoxes [9]	41.53	38.83	40.14
ResNet-TextBoxes [9] + Proposed	**45.20**	**46.92**	**46.12**

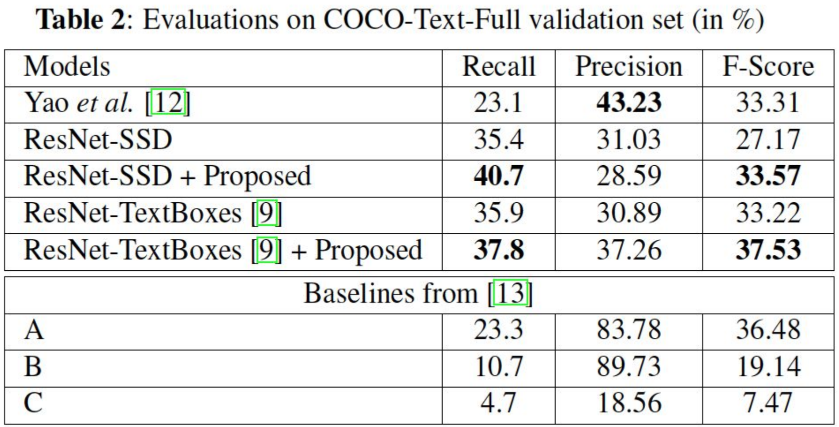
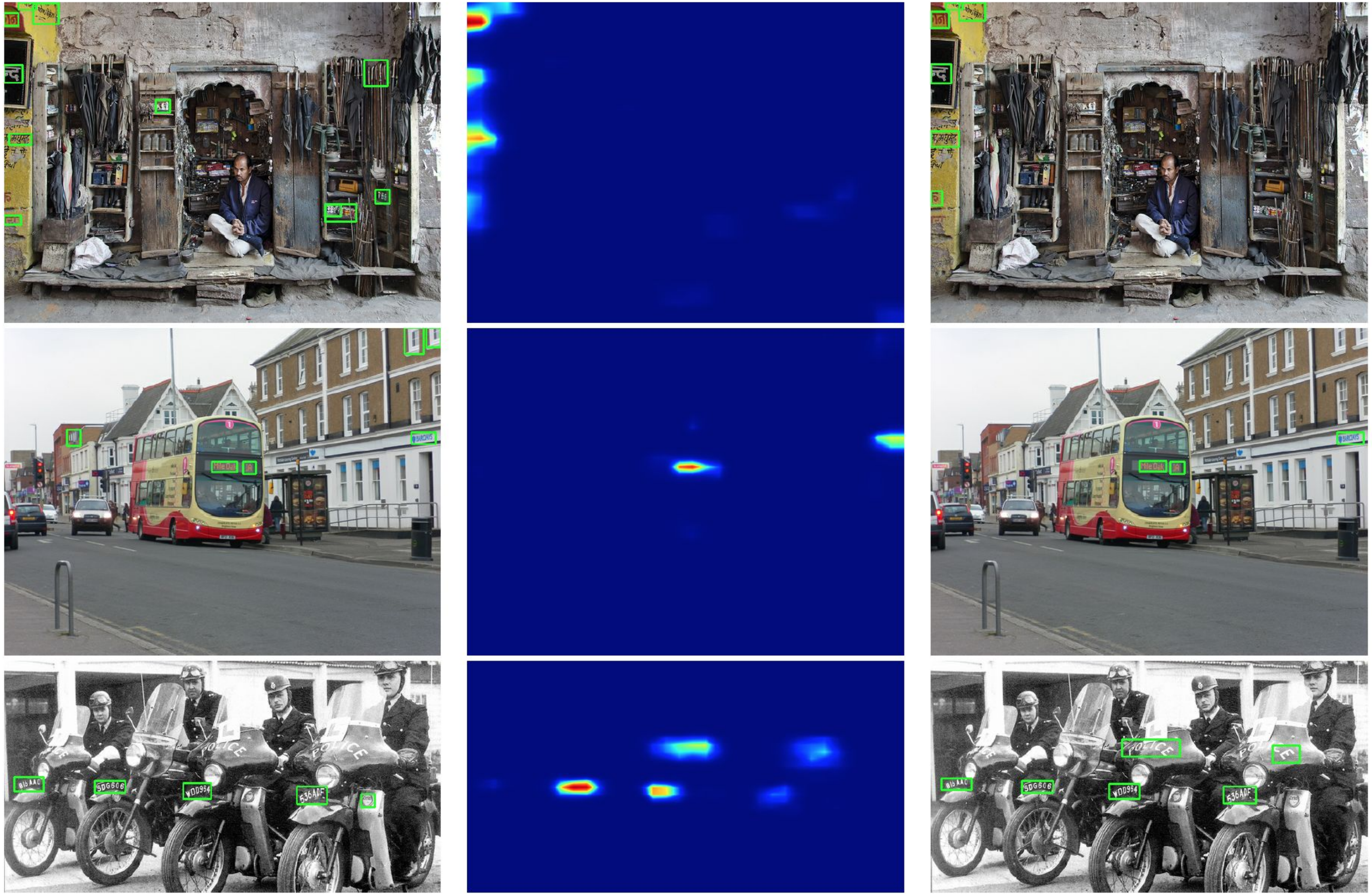


Table 2: Evaluations on COCO-Text-Full validation set (in %)

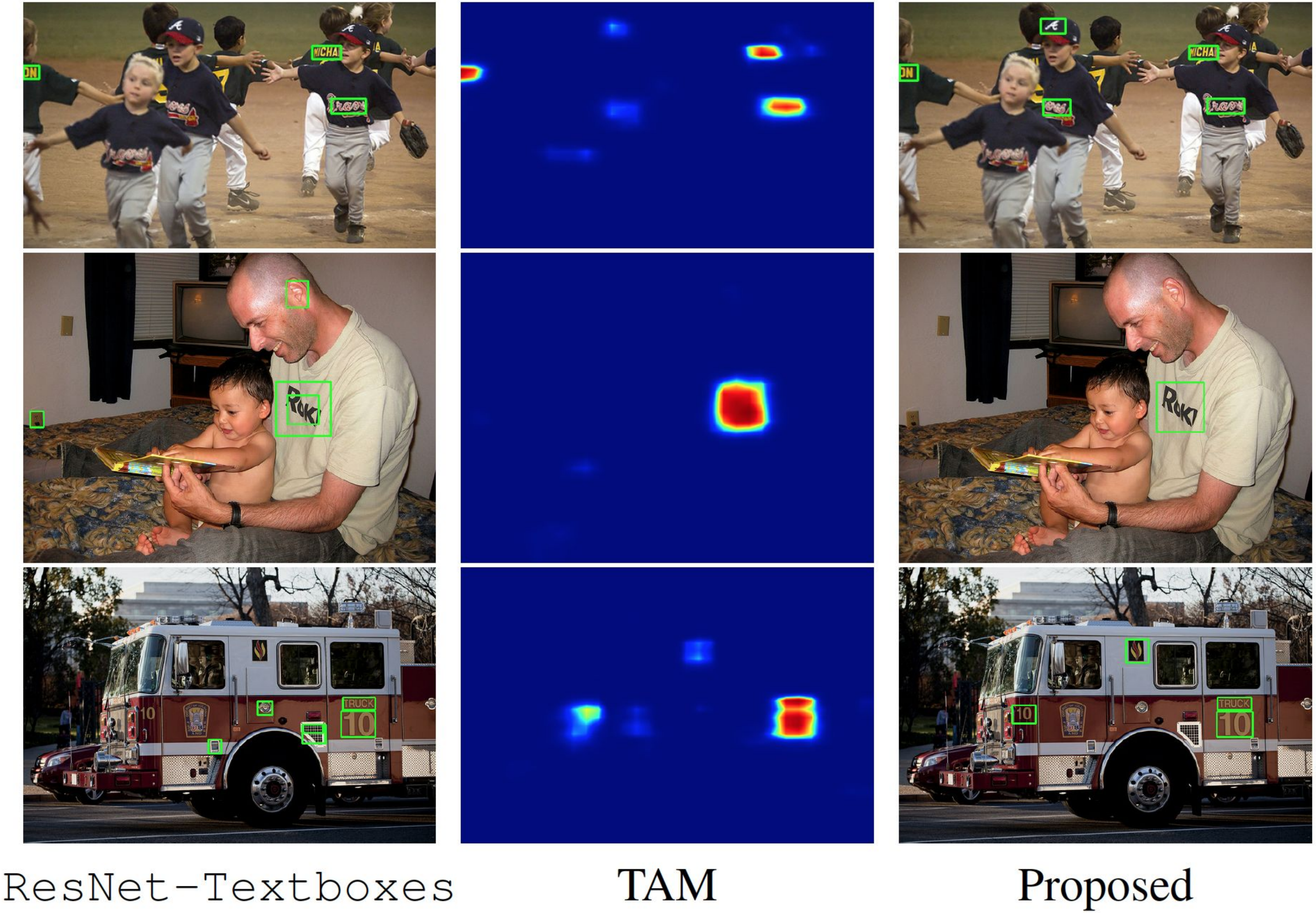
Models	Recall	Precision	F-Score
Yao *et al.* [12]	23.1	**43.23**	33.31
ResNet-SSD	35.4	31.03	27.17
ResNet-SSD + Proposed	**40.7**	28.59	**33.57**
ResNet-TextBoxes [9]	35.9	30.89	33.22
ResNet-TextBoxes [9] + Proposed	**37.8**	37.26	**37.53**
Baselines from [13]			
A	23.3	83.78	36.48
B	10.7	89.73	19.14
C	4.7	18.56	7.47

- Significant improvement over the baseline models.
- State-of-the-art performance on COCO-Text dataset, which is the most challenging text detection dateset up to date.

[9] Liao, Minghui, et al. "TextBoxes: A Fast Text Detector with a Single Deep Neural Network." *AAAI*. 2017.
[12] Yao, Cong, et al. "Scene text detection via holistic, multi-channel prediction." *arXiv preprint arXiv:1606.09002* (2016).



ResNet-SSD TAM Proposed



ResNet-Textboxes

TAM

Proposed

Conclusions and Future Work

- Main Take-aways:
  - Segmenter has a more global view
  - Detector is better at localizing accurate bounding boxes
  - Detector can be guided by the segmentation heatmap (TAM) via multiplicative gating on the feature maps
- Future Works:
  - Extend the proposed method to other detection task, eg. pedestrian detection
  - Optimize the proposed method for embedded systems

IEEE GlobalSIP

Thank you for your attention!
Any Questions?

Paper #1469: *Efficient Segmentation-Aided Text Detection for Intelligent Robots*

Junting Zhang*, Yuewei Na, Siyang Li, and C.-C. Jay Kuo.

*Email: juntingz@usc.edu

USC University of Southern California