SpotNet: Self-Attention Multi-Task Network for Object Detection

Hughes Perreault[1], Guillaume-Alexandre Bilodeau[1], Nicolas Saunier[1] and Maguelonne Héritier[2]

Polytechnique Montréal[1], Genetec[2]
Montréal, Canada

May 2020

POLYTECHNIQUE MONTRÉAL
TECHNOLOGICAL UNIVERSITY

Genetec

CRSNG NSERC

Motivation

- There is increasing interest in automatic **road user detection** for intelligent transportation systems, advanced driver assistance systems, traffic surveillance, etc.
- Inspired by the human visual attention, we aim to generate attention maps that can guide a network and improve the task of object detection.
- Given video sequences with bounding box ground-truth, we generate semi-supervised foreground/background annotations that can be used to train a segmentation module.
- The segmentation map thus produced is used inside the network as a self-attention mechanism to improve the object detection task.

Visual Attention

Visual Attention

Similarly to a human visual heatmap on an image, we train a network to produce an foreground/background segmentation map that acts as visual attention.

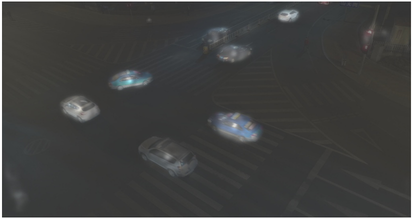

Figure 1: A visualisation of the attention map produced by SpotNet on top of its corresponding image, from the UAVDT [1] dataset.

Semi-Supervised Ground-Truth

- In order to train our visual attention, we need to produce segmentation ground-truth for non-annotated datasets.
- We produce such ground-truth with a background subtraction method (PAWCS [2]) for the fixed camera setting videos and with an optical flow method for the moving camera setting videos. We then intersect each imperfect segmentation mask with ground-truth bounding boxes in order to improve them.

Baseline

- We use CenterNet [3] as an object detection baseline upon which to build our model.
- CenterNet first processes an image through a backbone neural network. Using three heads, it then produces:
  - An object center heatmap.
  - A width and height for each point.
  - An offset for each point.

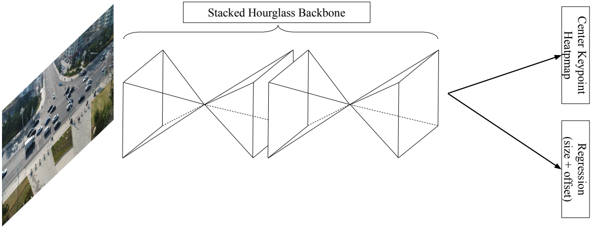

Figure 2: A representation of the CenterNet [3] model.

Self-Attention

- We improve upon the CenterNet model by implementing an internal attention mechanism, and train it using multi-task learning.
- We add a fourth head to the model, a foreground/background segmentation head, and train it using our semi-supervised ground-truth. The loss used here is the binary cross-entropy.
- The attention process works by multiplying each channel of the feature maps used by the other three branches by our attention map.

Complete Model

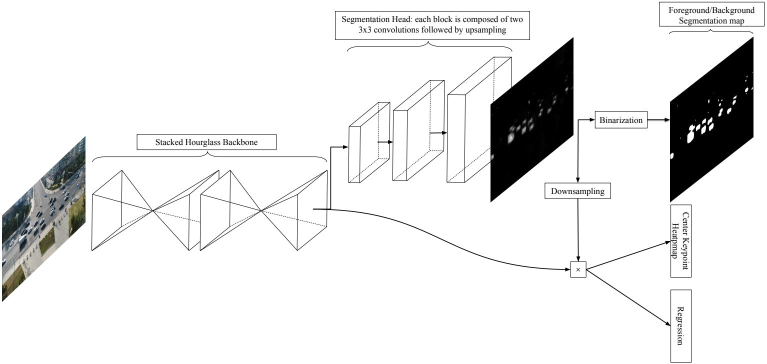

Figure 3: Overview of SpotNet: the input image first passes through a double-stacked hourglass network; the segmentation head then produces an attention map that multiplies the final feature map of the backbone network; the final center keypoint heatmap is then produced as well as the size and coordinate offset regressions for each object.

Datasets

- We trained and evaluated on two traffic surveillance datasets. UA-DETRAC [4] which has a fixed camera setting and UAVDT [1] which has moving camera setting.

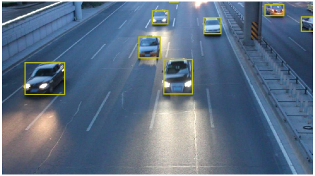

Figure 4: Sample from UA-DETRAC [4] with the ground-truth bounding boxes in yellow.

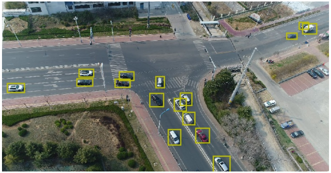

Figure 5: Sample from UAVDT [1] with the ground-truth bounding boxes in yellow.

Results on UA-DETRAC

Table 1: Results on the UA-DETRAC [4] dataset.

Model	Overall	Easy	Medium	Hard	Cloudy	Night	Rainy	Sunny
SpotNet (ours)	**86.80%**	**97.58%**	**92.57%**	**76.58%**	**89.38%**	**89.53%**	**80.93%**	**91.42%**
CenterNet[5]	83.48%	96.50%	90.15%	71.46%	85.01%	88.82%	77.78%	88.73%
FG-BR_Net [6]	79.96%	93.49%	83.60%	70.78%	87.36%	78.42%	70.50%	89.8%
HAT [7]	78.64%	93.44%	83.09%	68.04%	86.27%	78.00%	67.97%	88.78%
GP-FRCNNm [8]	77.96%	92.74%	82.39%	67.22%	83.23%	77.75%	70.17%	86.56%
R-FCN [9]	69.87%	93.32%	75.67%	54.31%	74.38%	75.09%	56.21%	84.08%
EB [10]	67.96%	89.65%	73.12%	53.64%	72.42%	73.93%	53.40%	83.73%
Faster R-CNN [11]	58.45%	82.75%	63.05%	44.25%	66.29%	69.85%	45.16%	62.34%
YOLOv2 [12]	57.72%	83.28%	62.25%	42.44%	57.97%	64.53%	47.84%	69.75%
RN-D [13]	54.69%	80.98%	59.13%	39.23%	59.88%	54.62%	41.11%	77.53%
3D-DETnet [14]	53.30%	66.66%	59.26%	43.22%	63.30%	52.90%	44.27%	71.26%

Results on UAVDT

Table 2: Results on the UAVDT [1] dataset.

Model	Overall
SpotNet (Ours)	**52.80%**
CenterNet[5]	51.18%
Wang *et al.* [15]	37.81%
R-FCN [9]	34.35%
SSD [16]	33.62%
Faster-RCNN [11]	22.32%
RON [17]	21.59%

Additional Results

Even though it is not our main goal, we evaluated the segmentation capabilities of our model on the Changedetection.net [18] dataset, and found out that we can outperform some classical methods but not the state-of-the-art.

Table 3: Results on the changedetection.net [18] dataset.

Model	Average F-Measure
PAWCS [2]	**0.872**
SuBSENSE [19]	0.831
SpotNet (Ours)	0.806
SGMM [20]	0.766
KNN [21]	0.731
GMM [22]	0.709

Additional Results

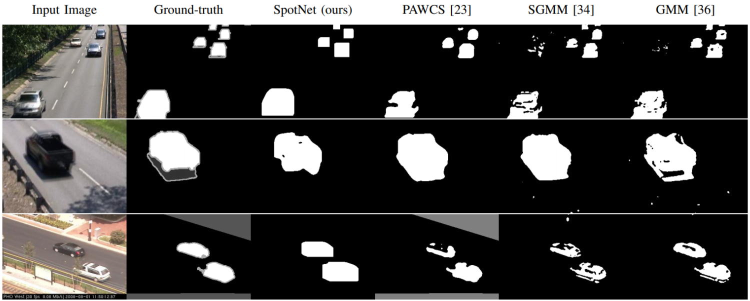

Figure 6: Example of foreground/background segmentation maps obtained with several segmentation methods. First row: frame 1015 of "highway", second row: frame 967 of "traffic", third row: frame 883 of "boulevard".

Ablation Study

Table 4: Ablation study on the UA-DETRAC [4] dataset.

Attention	Multi-Task	Overall	Easy	Medium	Hard	Cloudy	Night	Rainy	Sunny
✓	✓	**86.80%**	**97.58%**	**92.57%**	**76.58%**	**89.38%**	**89.53%**	**80.93%**	**91.42%**
	✓	84.57%	96.72%	90.85%	73.16%	86.53%	88.76%	78.84%	90.10%
		83.48%	96.50%	90.15%	71.46%	85.01%	88.82%	77.78%	88.73%

Limitations

- Our model needs semi-supervised annotations to be trained properly.
- However, we believe that in most real-world applications, sequences are available and we can thus run background subtraction or optical flow to generate them.

Conclusion

- We presented a novel multi-task model equipped with a self-attention process.
- We trained it with semi-supervised annotations and a multi-task loss.
- We show that these improvements allow us to reach state-of-the-art performance on two traffic scene datasets with different settings.
- We argue that not only does this improve accuracy by a large margin, it also provides instance segmentations of the road users almost at no cost.

Acknowledgment

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [RDCPJ 508883 - 17], and the support of Genetec.

References I

D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 370–386, 2018.

P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "A self-adjusting approach to change detection based on background word consensus," in *2015 IEEE winter conference on applications of computer vision*, pp. 990–997, IEEE, 2015.

X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

L. Wen, D. Du, Z. Cai, Z. Lei, M.-C. Chang, H. Qi, J. Lim, M.-H. Yang, and S. Lyu, " UA-DETRAC: A New Benchmark and Protocol for Multi-Object Detection and Tracking," *arXiv CoRR*, vol. abs/1511.04136, 2015.

K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6569–6578, 2019.

Z. Fu, Y. Chen, H. Yong, R. Jiang, L. Zhang, and X.-S. Hua, "Foreground gating and background refining network for surveillance object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 6077–6090, 2019.

S. Wu, M. Kan, S. Shan, and X. Chen, "Hierarchical attention for part-aware face detection," *International Journal of Computer Vision*, vol. 127, no. 6-7, pp. 560–578, 2019.

S. Amin and F. Galasso, "Geometric proposals for faster r-cnn," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, IEEE, 2017.

References II

- J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems 29*, pp. 379–387, Curran Associates, Inc., 2016.
- L. Wang, Y. Lu, H. Wang, Y. Zheng, H. Ye, and X. Xue, "Evolving boxes for fast vehicle detection," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1135–1140, IEEE, 2017.
- S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.
- J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- H. Perreault, G.-A. Bilodeau, N. Saunier, and P. Gravel, "Road user detection in videos," *arXiv preprint arXiv:1903.12049*, 2019.
- S. Li and F. Chen, "3d-detnet: a single stage video-based vehicle detector," in *Third International Workshop on Pattern Recognition*, vol. 10828, p. 108280A, International Society for Optics and Photonics, 2018.
- T. Wang, R. M. Anwer, H. Cholakkal, F. S. Khan, Y. Pang, and L. Shao, "Learning rich features at high-speed for single-shot object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1971–1980, 2019.
- W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, 2016.

References III

T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen, "Ron: Reverse connection with objectness prior networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, p. 2, 2017.

N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection. net: A new change detection benchmark dataset," in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pp. 1–8, IEEE, 2012.

P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "Subsense: A universal change detection method with local adaptive sensitivity," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 359–373, 2014.

R. H. Evangelio, M. Pätzold, and T. Sikora, "Splitting gaussians in mixture models," in *2012 IEEE Ninth international conference on advanced video and signal-based surveillance*, pp. 300–305, IEEE, 2012.

Z. Zivkovic and F. Van Der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern recognition letters*, vol. 27, no. 7, pp. 773–780, 2006.

C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 2, pp. 246–252, IEEE, 1999.