POLYTECHNIQUE MONTRÉAL
TECHNOLOGICAL UNIVERSITY

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

Abstract

In this paper we train an object detection network to produce a foreground/background segmentation map as well as bounding boxes via a multi-task learning approach, and we use this map in a self-attention mechanism. To train the segmentation map, we produce semi-supervised ground-truth using background subtraction or optical flow. We show that by using this method, we obtain a significant mAP improvement on two traffic surveillance datasets, with state-of-the-art results on both UA-DETRAC and UAVDT.

Project summary

There is increasing interest in automatic road user detection for intelligent transportation systems, advanced driver assistance systems, traffic surveillance, etc. Given video sequences with bounding box ground-truth, we aim to generate semi-supervised foreground/background annotations that can be used to train a segmentation head. The segmentation map, visualised in figure 3, is used inside the network as a self-attention mechanism to improve the object detection task.

Baseline: Centernet [1]

- We use CenterNet [1] as a 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.

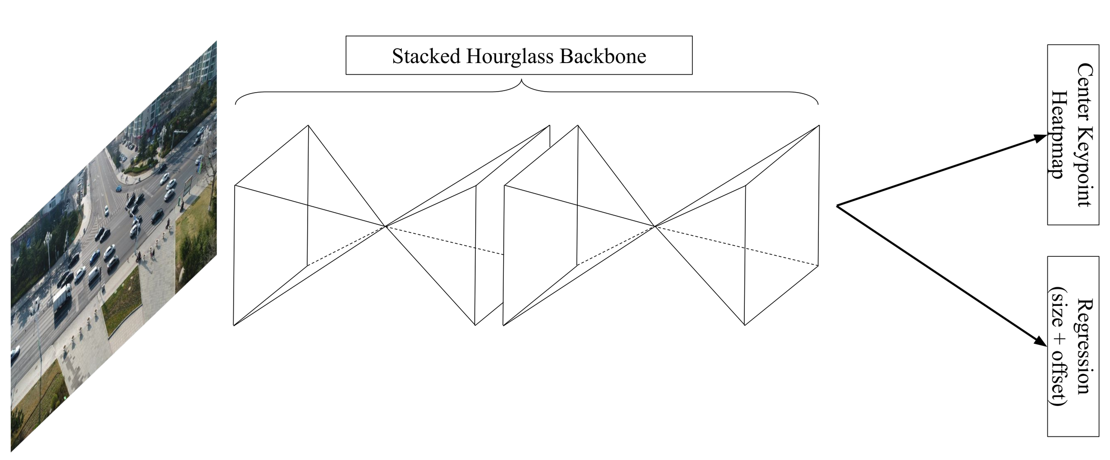

Figure 1:A representation of the CenterNet [1] model.

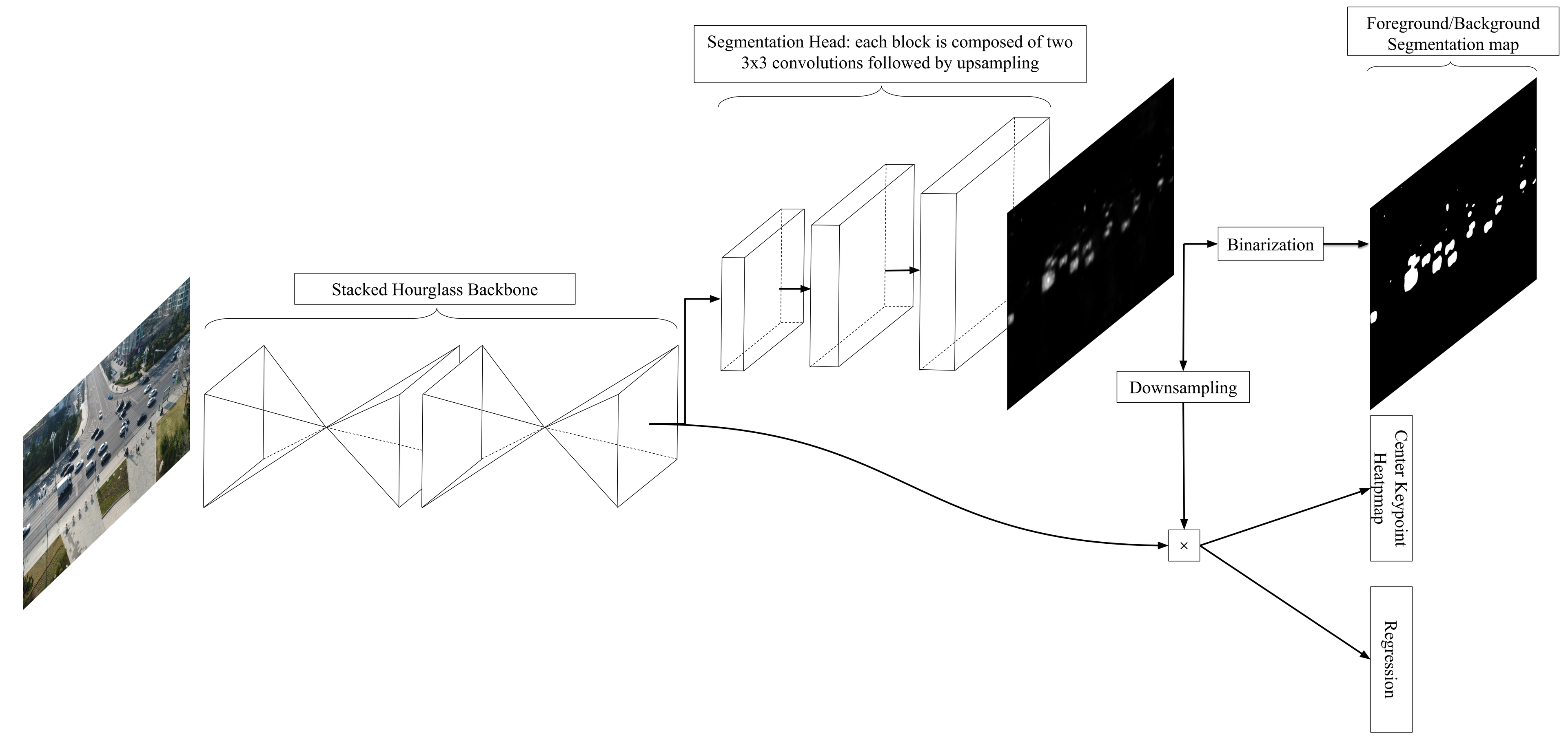

Figure 2:Overview of SpotNet: the input image first passes through a double-stacked hourglass network; the segmentation head then produces an attention map that multiplies the final feature map of the backbone network; the final center keypoint heatmap is then produced as well as the size and coordinate offset regressions for each object.

Self-Attention

We improve upon the CenterNet model (figure 1) 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 obtained with background subtraction and optical flow (figure 2). 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.

Results on UA-DETRAC [2]

Table 1:Results on the UA-DETRAC [2] 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[3]	83.48%	96.50%	90.15%	71.46%	85.01%	88.82%	77.78%	88.73%
FG-BR_Net	79.96%	93.49%	83.60%	70.78%	87.36%	78.42%	70.50%	89.8%
HAT	78.64%	93.44%	83.09%	68.04%	86.27%	78.00%	67.97%	88.78%
GP-FRCNNm	77.96%	92.74%	82.39%	67.22%	83.23%	77.75%	70.17%	86.56%
R-FCN	69.87%	93.32%	75.67%	54.31%	74.38%	75.09%	56.21%	84.08%
EB	67.96%	89.65%	73.12%	53.64%	72.42%	73.93%	53.40%	83.73%
Faster R-CNN	58.45%	82.75%	63.05%	44.25%	66.29%	69.85%	45.16%	62.34%
YOLOv2	57.72%	83.28%	62.25%	42.44%	57.97%	64.53%	47.84%	69.75%
RN-D	54.69%	80.98%	59.13%	39.23%	59.88%	54.62%	41.11%	77.53%
3D-DETnet	53.30%	66.66%	59.26%	43.22%	63.30%	52.90%	44.27%	71.26%

Results on UAVDT [4]

Table 2:Results on the UAVDT [4] dataset.

Model	Overall
SpotNet (Ours)	**52.80%**
CenterNet[3]	51.18%
Wang *et al.* [5]	37.81%
R-FCN	34.35%
SSD	33.62%
Faster-RCNN	22.32%
RON	21.59%

Additional results

Even tough it is not our main goal, we evaluated the segmentation capabilities of our model on the Changedetection.net [6] 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 [6] dataset.

Model	Average F-Measure
PAWCS	**0.872**
SuBSENSE	0.831
SpotNet (Ours)	0.806
SGMM	0.766
KNN	0.731
GMM	0.709

Visual Attention

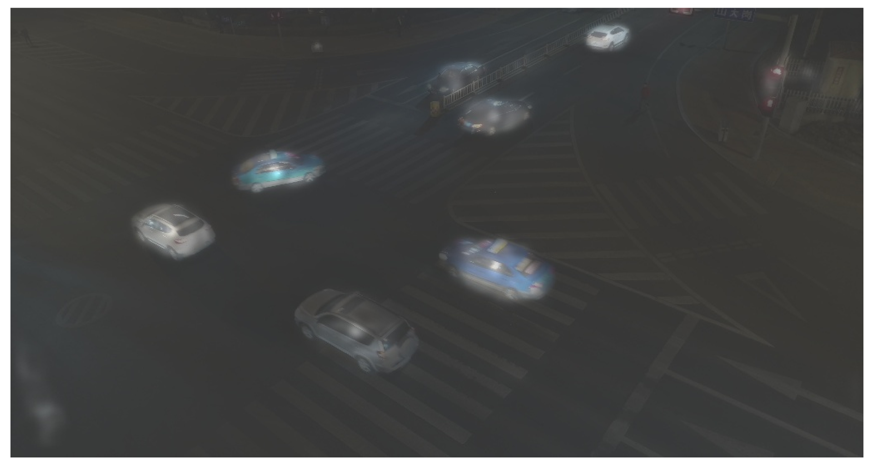

Figure 3:A visualisation of the attention map produced by SpotNet on top of its corresponding image, from the UAVDT [4] dataset.

Conclusion

- We presented a novel multi-task model equipped with a self-attention process.
- We trained it with semi-supervised annotations and 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.

References

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

[2] 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.

[3] 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*, 2019, pp. 6569–6578.

[4] 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)*, 2018, pp. 370–386.

[5] 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*, 2019, pp. 1971–1980.

[6] 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.* IEEE, 2012, pp. 1–8.

Acknowledgements

Genetec
CRSNG NSERC