Instance Segmentation as Image Segmentation Annotation

Thomio Watanabe and Denis F. Wolf

Mobile Robotics Laboratory, Institute of Mathematics and Computer Science
University of São Paulo, São Carlos, Brazil

Abstract

This work approaches the instance segmentation as an annotation problem. It extends our previous work, the DCME, to provide a solution more computational efficient and improve evaluation scores. We solve the instance segmentation with a single segmentation network and we make DCME robust to different object sizes.

Introduction

The two main approaches to solve the instance segmentation problem are:

- Network problem
  - Multiple networks or multi-task network
  - High scores but high computational cost
- Annotation problem
  - Single segmentation network
  - Low scores but low computational cost

Multi-task networks directly specify sub-tasks and this may be too restrictive for deep learning.

Encoding

The Distance to Center of Mass Encoding (DCME), Watanabe and Wolf [1], was able to generate class-agnostic instance masks.

- Associates a superficial center of mass (CM) to each object instance
- Pixels that belong to an instance present displacement vectors pointing to its CM

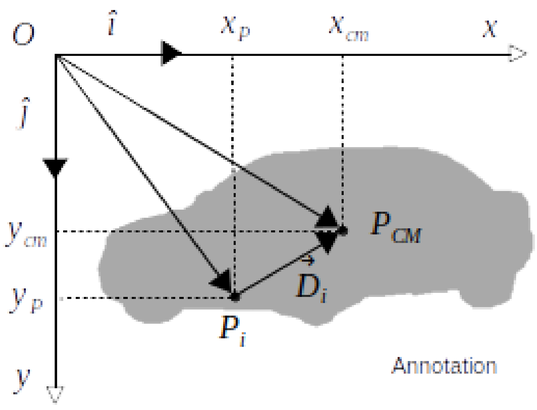

Figure: DCME encoding.

Objectives

Extend the DCME to:

- Solve the instance segmentation problem with a single segmentation network
  - Re-purpose network encoder to classify instances
- Improve evaluation scores (mean AP)
  - Loss function robust to long displacement vectors

Instance Classification

Reusing classification networks in detection networks is a bad idea:

- Classification problem: single object centralized in image
- Detection problem: multiple objects anywhere in image

We repurpose the encoder to perform a segmentation. We create an image grid and we use the encoder to classify each block. For n $(2, 2)$ stride operations the grid size is given by G_s and it is the same for both horizontal and vertical dimensions.

$$G_s = 2^n \tag{1}$$

During the training process the class labels are defined according to Equation 2, where $P(x, y)_{image}$ is the image pixel position and $P(x, y)_{encoder}$ is the encoder output position.

$$P(x, y)_{encoder} = floor\left(\frac{P(x, y)_{image}}{G_s}\right) \tag{2}$$

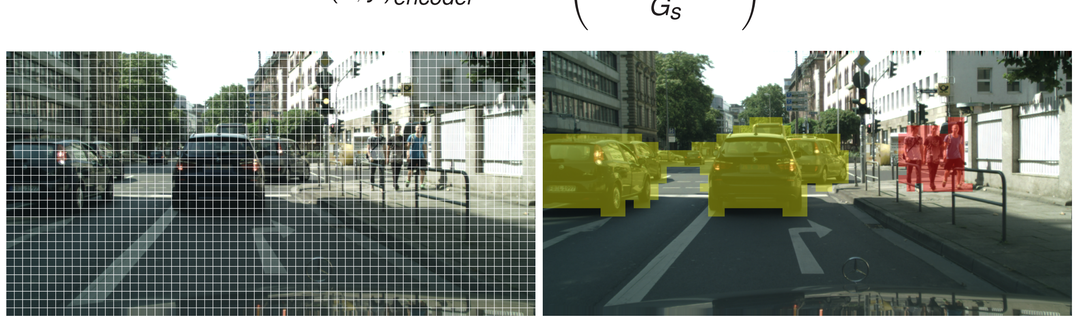

Figure: **Left:** input image with (512,1024) resolution and grid size of 16. Presents 32 vertical blocks and 64 horizontal blocks. **Right:** Input image blocks classification. Car class in yellow and people class in red.

Loss Function

Once the DCME encoding is based on 2D displacement vectors, errors in bigger objects will generate very high error values while errors in small objects will be insignificant. This behavior generate biased models that will preferably detect large objects.

Independent outputs

Every single output pixel from both output channels in the network decoder were considered as an independent output. The model general loss was evaluated considering this, with the number of samples defined by 2 channels, the number of images n and the decoder spatial dimensions (r, c).

$$N = 2 \cdot n \cdot r \cdot c \tag{3}$$

Gradients amplitude

The gradients were clipped with a translated version of the logistic function, Equation 4. In this approach A is a parameter that must be adjusted according to the input image resolution.

$$f(x) = \frac{A}{2} \cdot \left(\frac{e^x - 1}{e^x + 1}\right) \tag{4}$$

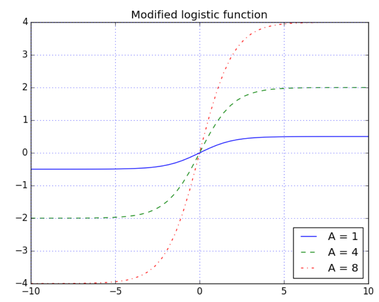

Figure: Translated logistic function used to clip gradients.

Results

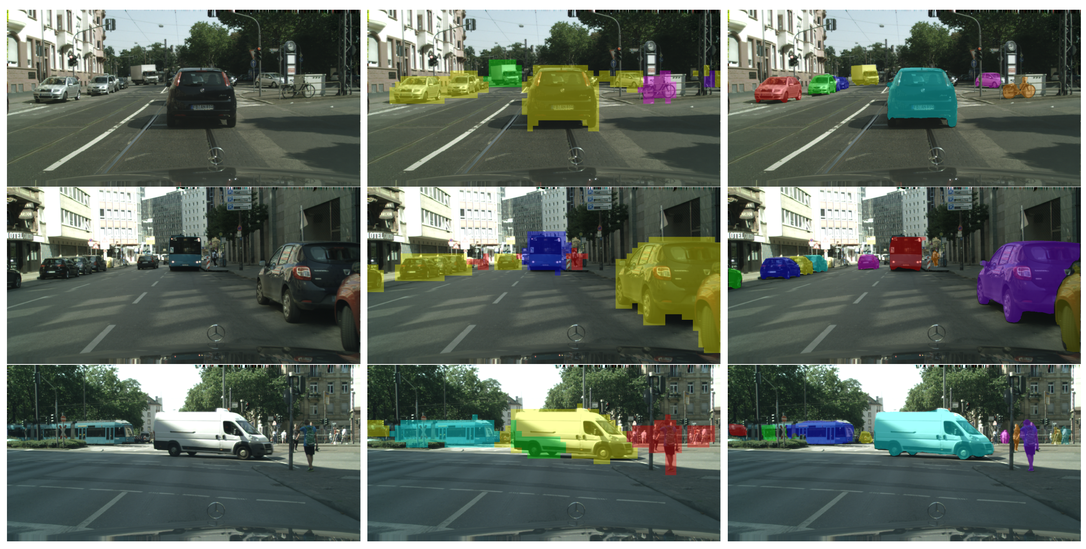

Figure: Input images, class map and instance masks on cityscapes validation set.

	AP		
Class	Ours	DCME	Multitask
person	6.66	1.77	19.22
rider	3.09	0.71	21.39
car	24.14	15.53	36.57
truck	6.02	2.00	18.80
bus	9.76	4.30	26.82
train	6.41	4.57	15.88
motorcycle	3.62	0.93	19.39
bicycle	2.08	0.33	14.51
mean	7.72	3.77	21.57

Table: Ours, DCME [1] and Multitask Learning [2] per class evaluation results on Cityscapes test set.

Conclusions

- We solve the instance segmentation problem with a single segmentation network.
- Compared to multi-task networks it presents a lower computational cost.
- This solution solves the partial occlusion problem.

References

[1] T. Watanabe and D. Wolf, "Distance to center of mass encoding for instance segmentation," in *The 21st IEEE International Conference on Intelligent Transportation Systems (ITSC)*, Hawaii, 2018.
[2] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Acknowledgments

This research was funded by Sao Paulo Research Foundation (FAPESP) project: #2015/26293-0.