Boundary-Aware Instance Segmentation

Zeeshan Hayder[1,2], Xuming He[2,1], Mathieu Salzmann[3]

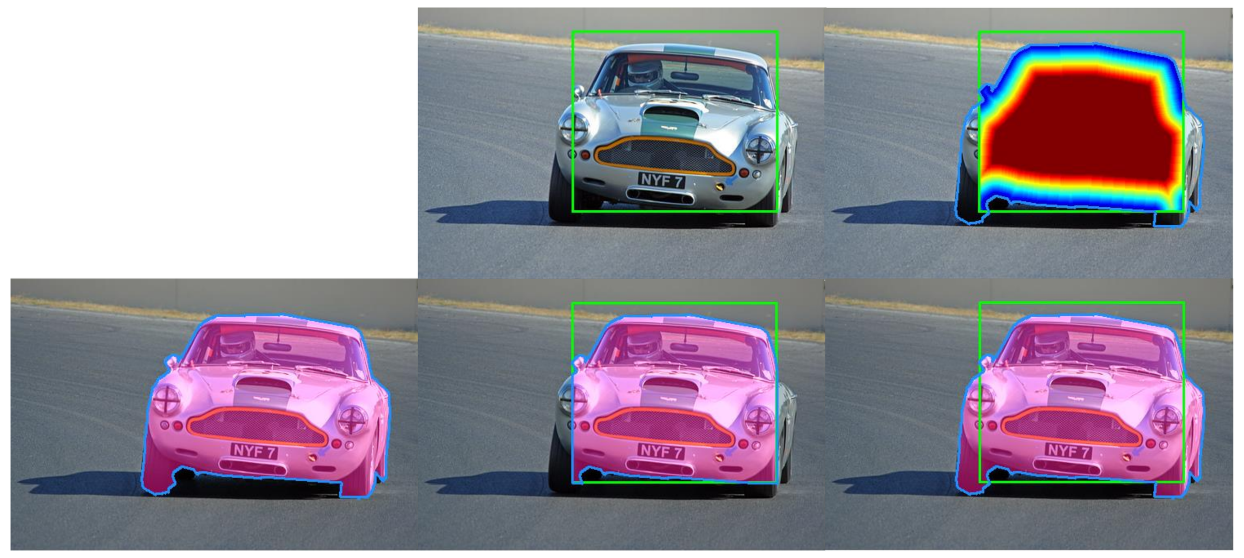
[1]Australian National University, [2]Data 61 / CSIRO, Australia, [3]CVLab, EPFL, Switzerland

{zeeshan.hayder, xuming.he}@anu.edu.au, mathieu.salzmann@epfl.ch

Goals and Contributions

Instance-level semantic segmentation (IS)

- Key component of scene understanding pipelines.
- Jointly detect, segment and classify all individual object instances.
- Provides detailed information about the location, shape and number of individual objects.



Input/ground-truth Traditional methods Our approach

Existing IS methods

- Existing methods typically follow a detect-then-segment approach [1, 3, 7-9].
- They limit the mask prediction to bounding-boxes and are thus sensitive to the boxes quality.

Our contribution: Boundary-Aware IS

- Predict masks beyond the scope of bounding boxes.
- Novel segment representation based on the distance transform of object masks.
- Fully-differential object mask network (OMN) with residual-deconvolutional architecture.
- End-to-end learning using Boundary-aware Instance Segmentation Network (BAIS) based on MNC [5].

Mask Representation

- Each bounding-box depicts a partially-observed object instance.
- Bounding-boxes are warped to a fixed size to encode a scale invariant shape representation.
- At each inside pixel p, ground-truth mask represents minimum distance to the true object boundary Q.
- Truncate the distance transform to obtain a restricted range of values $[0, \cdots, R]$

$$D(p) = \min\left(\min_{\forall q \in Q} \lceil d(p,q) \rceil, R\right) ,$$

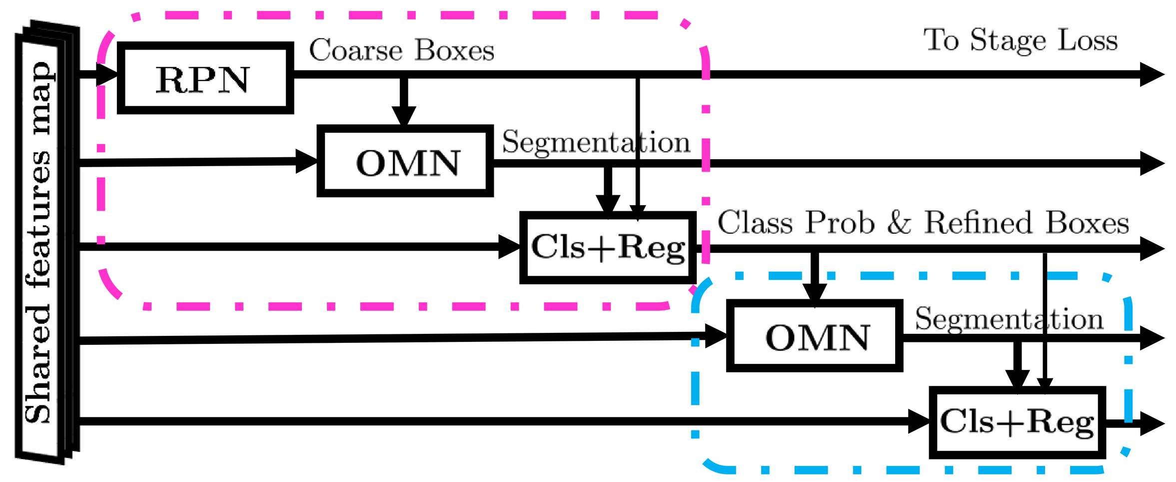
- One-hot encoding by quantizing distance map $D(p)$ into K-dimensional binary vector $b(p)$

$$D(p) = \sum_{n=1}^{K} r_n \cdot b_n(p), \quad \sum_{n=1}^{K} b_n(p) = 1 ,$$

- Efficient decoding by taking the union of all the disks $T(p,r)$ of radius r at pixel p

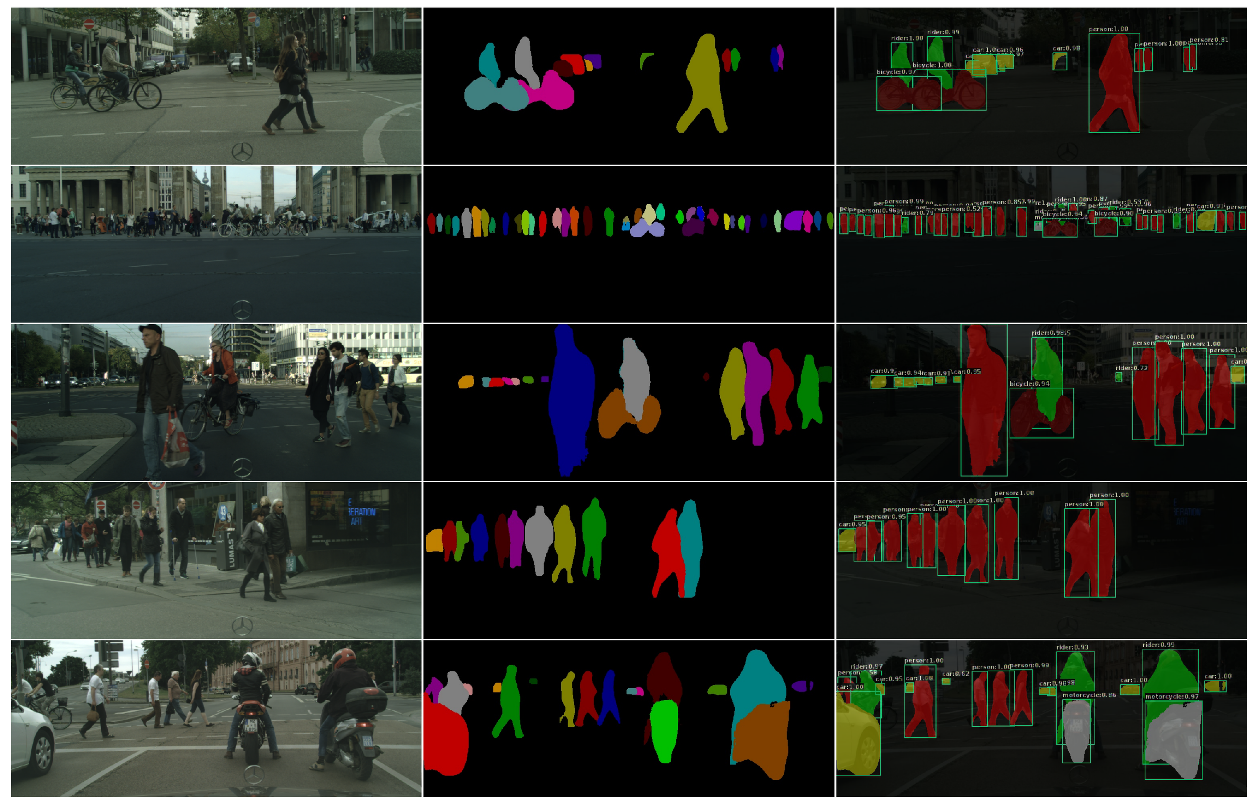
$$M = \bigcup_{n=1}^{K} \bigcup_{p} T(p, r_n) = \bigcup_{n=1}^{K} T(\cdot, r_n) * B_n .$$

Multi-stage Multi-task Learning

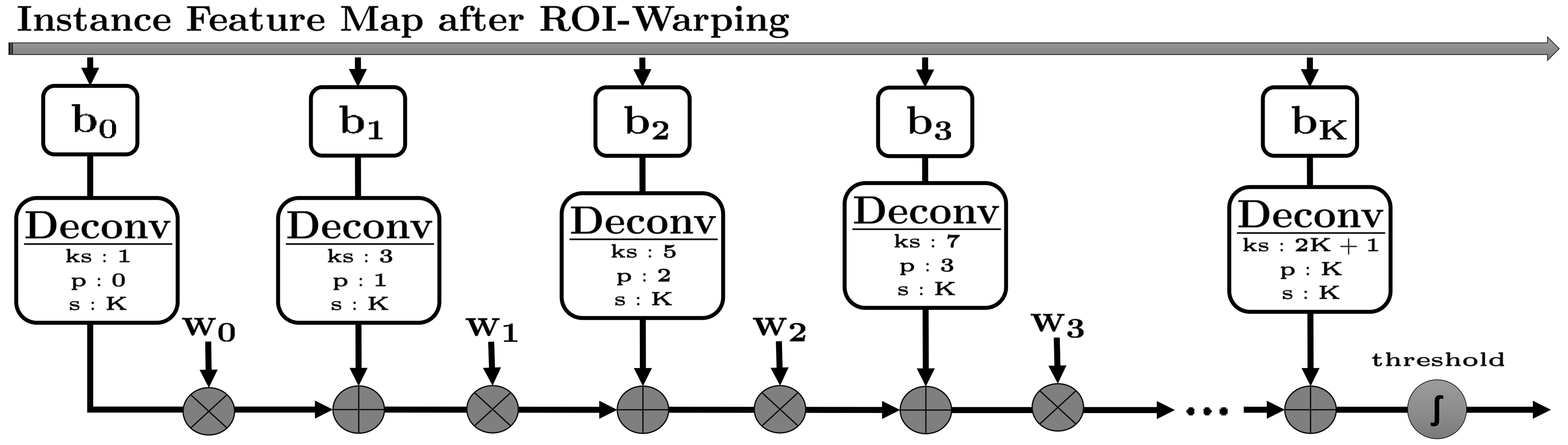


End-to-end learning via OMN & Multitask Network Cascade [5]

Qualitative Results



Residual-deconvolutional Subnet for Decoding



Quantitative Evaluation

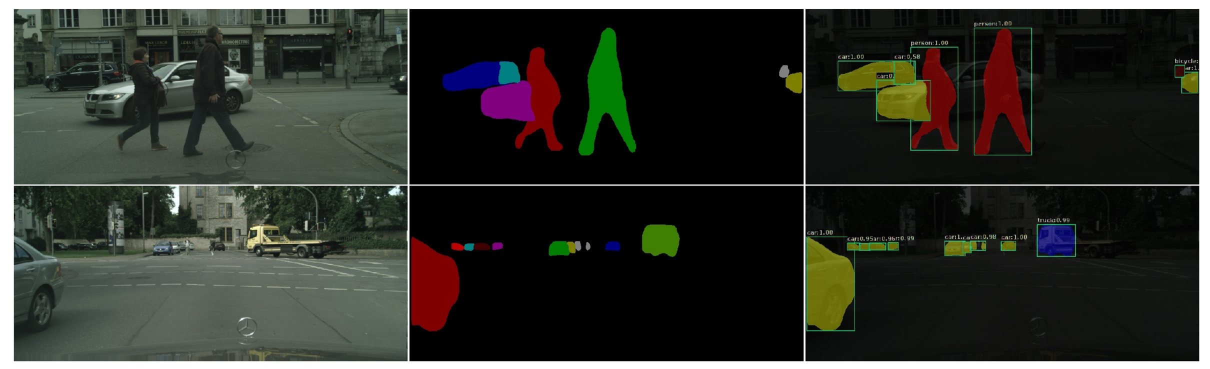
- **Pascal 2012 val dataset [12]**

VOC 2012 (val)	mAP (0.5)	mAP (0.7)	time/img (s)
SDS [1]	49.7	25.3	48
PFN [2]	58.7	42.5	~ 1
Hypercolumn [3]	60.0	40.4	>80
InstanceFCN [4]	61.5	43.0	1.50
MNC [5]	63.5	41.5	0.36
MNC-new [5]	65.01	46.23	0.42
BAIS - insideBBox (ours)	64.97	44.58	0.75
BAIS - full (ours)	**65.69**	**48.30**	0.78

- **Cityscapes dataset [7]**

Cityscapes (test)	AP	AP (50%)	AP (100m)	AP (50m)
ILSVDC [6]	2.3	3.7	3.9	4.9
MCG+RCNN [7]	4.6	12.9	7.7	10.3
PEIS [8]	8.9	21.1	15.3	16.7
RecAttend [9]	9.5	18.9	16.8	20.9
InstanceCut [10]	13.0	27.9	22.1	26.1
BAIS - full (ours)	17.4	**36.7**	29.3	34.0
DWT + PSPNet [11]	**19.4**	35.3	**31.4**	**36.8**

Failure Cases



[1] B. Hariharan, et al. Simultaneous Detection and Segmentation. In ECCV, 2014.
[2] X. Liang, et al. Proposal-free Network for Instance-level Object Segmentation. In CoRR, 2015.
[3] B. Hariharan, et al. Hypercolumns for object segmentation and fine-grained localization. In CVPR, 2015.
[4] J. Dai, et al. Instance-sensitive Fully Convolutional Networks. ECCV, 2016.
[5] J. Dai, et al. Instance-aware Semantic Segmentation via Multi-task Network Cascades. In CVPR, 2016.
[6] J. Van, et al. Instance-level Segmentation of Vehicles using Deep Contours. In ACCV, 2016.
[7] M. Cordts, et al. The Cityscapes Dataset for Semantic Urban Scene Understanding. In CVPR, 2016.
[8] J. Uhrig, et al. Pixel-Level Encoding and Depth Layering for Instance Semantic Labeling. In GCPR, 2016.
[9] M. Ren, et al. End-to-End Instance Segmentation and Counting with Recurrent Attention. In CVPR, 2017.
[10] A. Kirillov, et al. Instance-Cut: from Edges to Instances with Multicut. In CVPR, 2017.
[11] M. Bai, et al. Deep Watershed Transform for Instance Segmentation. In CVPR, 2017.
[12] M. Everingham, et al. The Pascal Visual Object Classes (VOC) Challenge. In IJCV, 2010.