GSTO: Gated Scale-Transfer Operation for Multi-Scale Feature Learning in Semantic Segmentation

Zhuoying Wang, Yongtao Wang*, Zhi Tang, Yangyan, Ying Chen, Haibin Ling and Weisi Lin

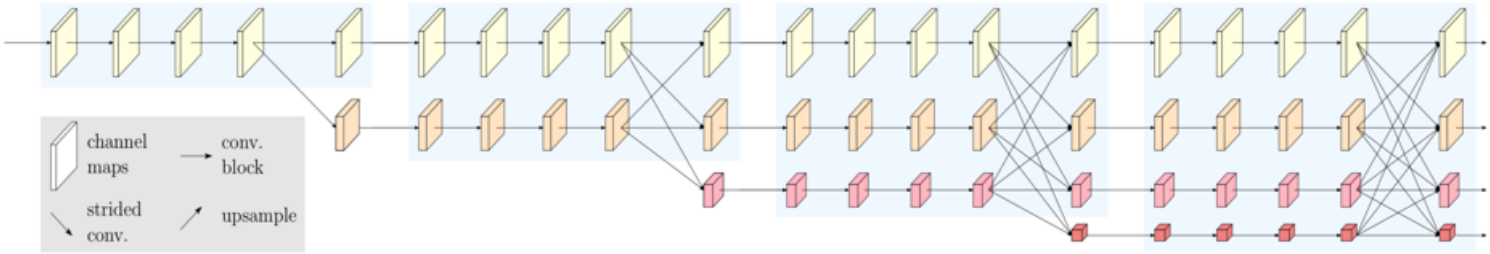
Wangxuan Institute of Computer Technology, Peking University
Email: {wzypku, wyt, tangzhi}@pku.edu.cn, {yangyan.lyy, chenying.ailab}@alibaba-inc.com, hling@cs.stonybrook.edu, wslin@ntu.edu.sg

January 10, 2020

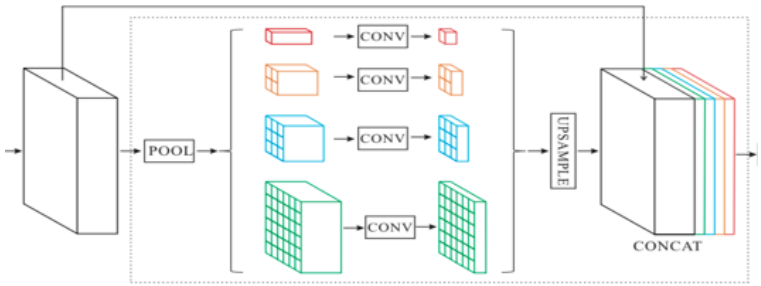
Outline

Multi-Scale Learning

Existing CNN-based methods for pixel labeling heavily depend on multi-scale features to meet the requirements of both semantic comprehension and detail preservation.



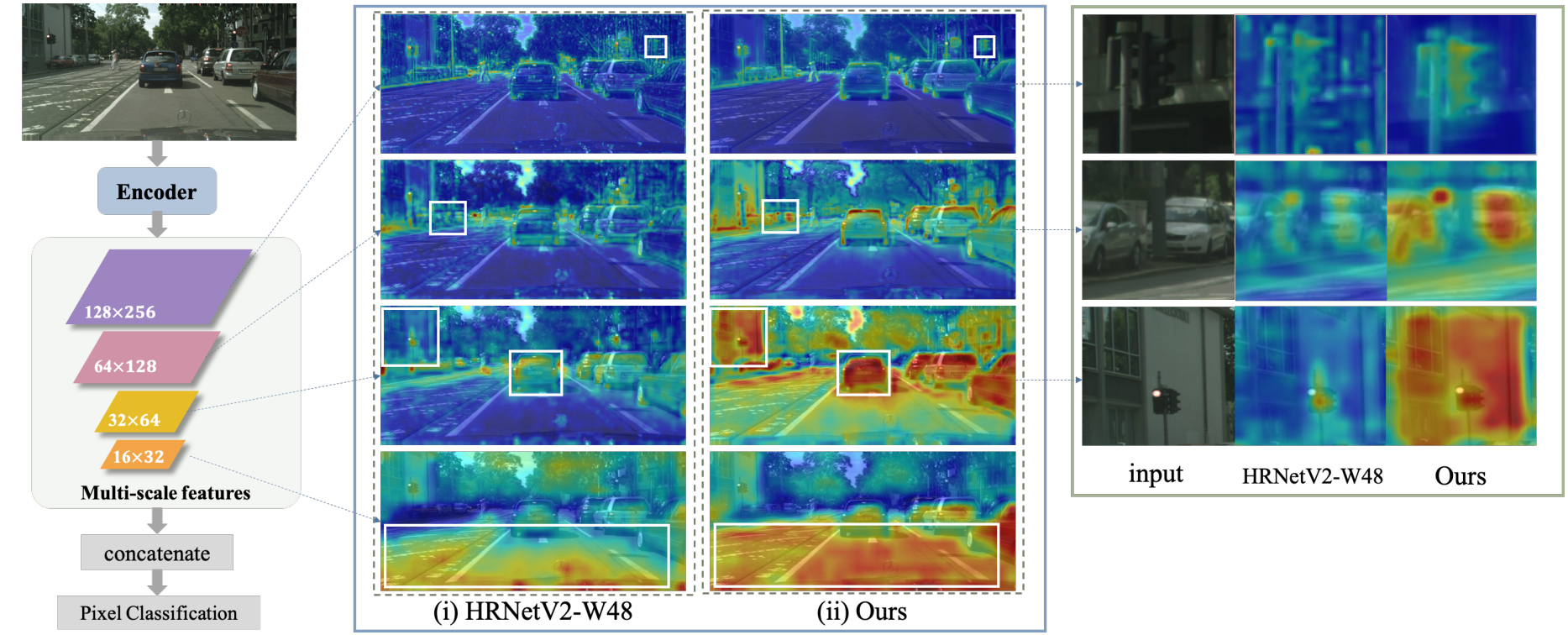
(a)



(b)

Figure: (a) High-Resolution Network. (b) Pyramid Pooling Module.

Scale Confusion



(a) Multi-scale architecture (b) Multi-scale features (c) Detailed presentation

Figure: Visual comparison of the multi-scale features extracted by the encoder of (i) HRNetV2-W48 and (ii) our proposed GSTO-HRNet. Each heat map is obtained by averaging the corresponding feature map along the channel dimension, and warmer color (red) indicates larger activation.

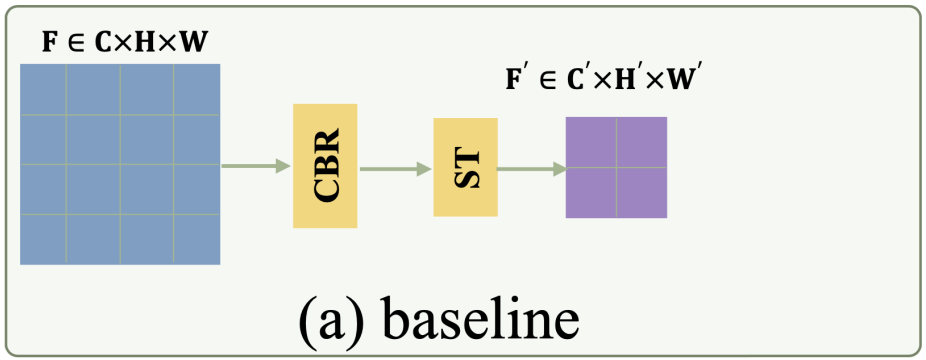
Outline

Traditional Scale-Transfer Operation

traditional transition is performed through down-sampling like average pooling and up-sampling like bilinear interpolation.

$$\widetilde{F}_{kij} = \sum_{m=1}^{C} \omega_{km} \cdot F_{mij}, \qquad k = 1, ..., C', \tag{1}$$

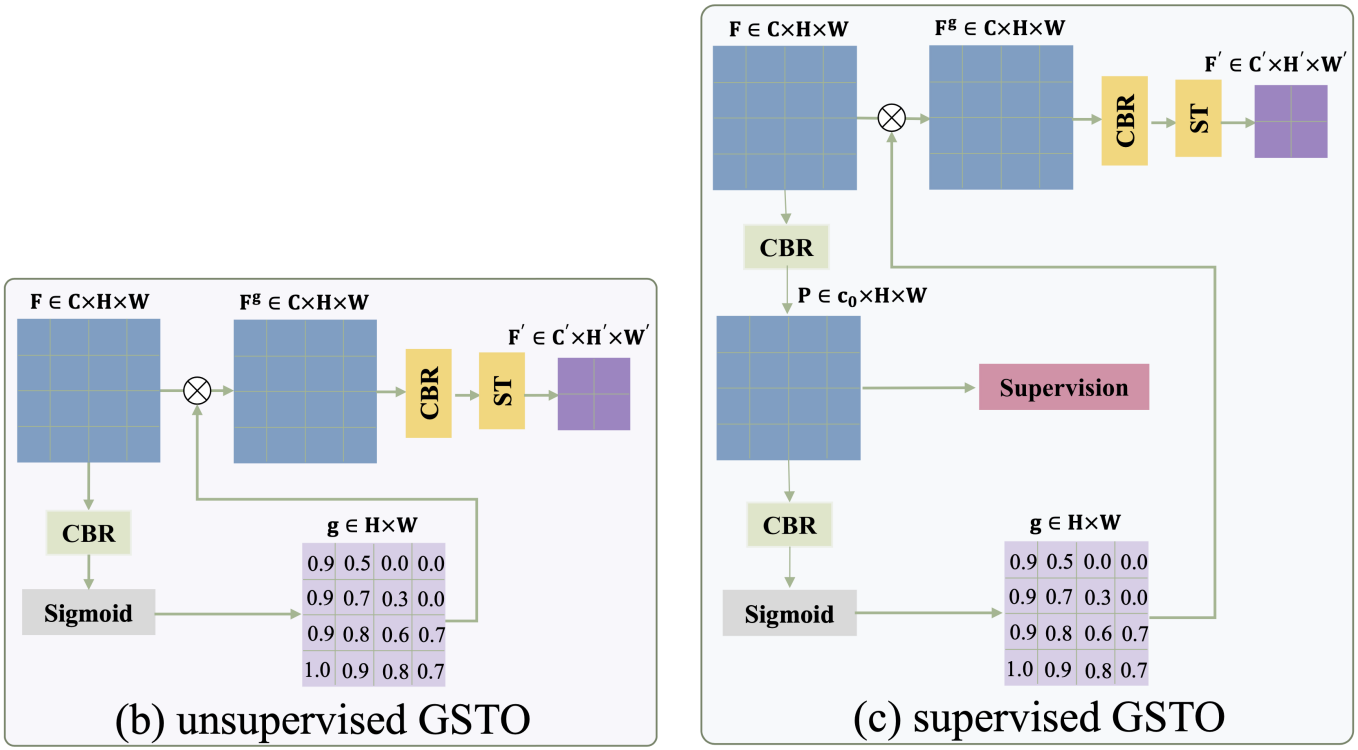
$$F' = ST(\widetilde{F}), \tag{2}$$



Gated Scale-Transfer Operation

In the proposed GSTOs, a spatially gated feature F^g is produced first and then Equations 1 and 2 are performed on F^g instead of on the original F

$$F^g_{mij} = g_{ij} \cdot F_{mij}, \qquad m = 1, ..., C, \tag{3}$$

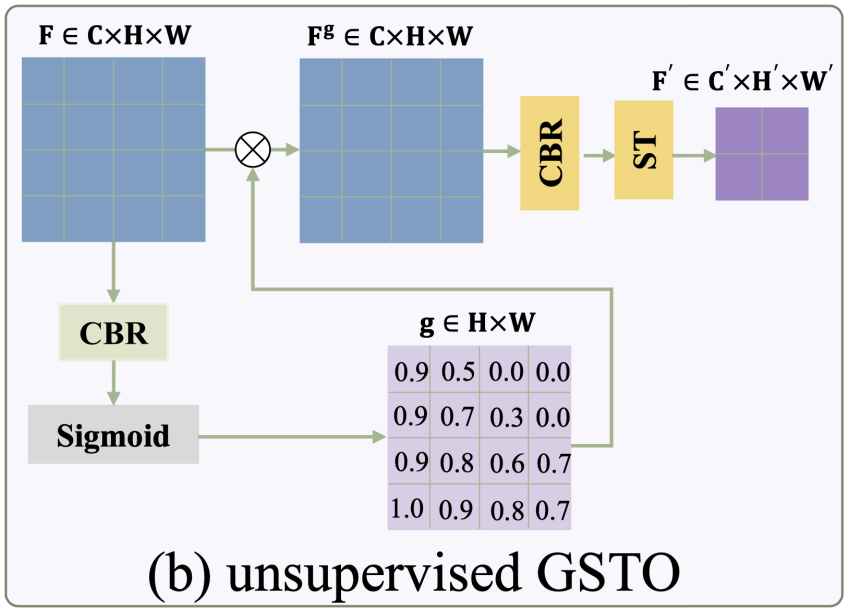


Gated Scale-Transfer Operation

unsupervised GSTO

The element of the gate g_{ij} is calculated from the original feature F, by an 1×1 convolution with input channel of C and output channel of 1, followed by sigmoid.

$$g_{ij} = \sigma(\sum_{m=1}^{C} \rho_m \cdot F_{mij}) \tag{4}$$



(b) unsupervised GSTO

Gated Scale-Transfer Operation

supervised GSTO

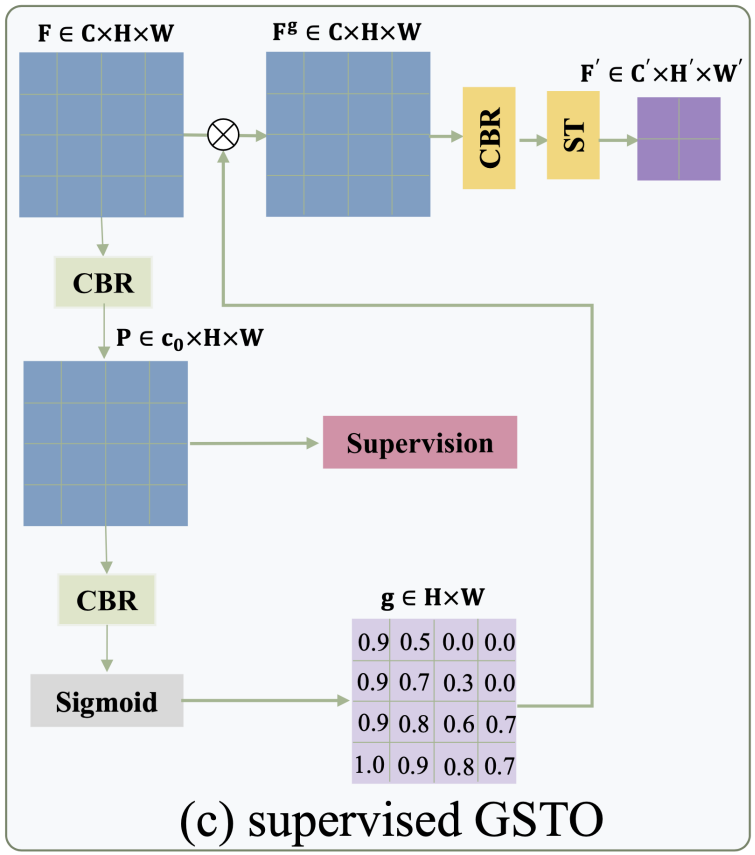
A light-weight predictor, such as a 1×1 convolution, is performed on F to get $P \in \mathbb{R}^{c_0 \times H \times W}$, where c_0 is the number of semantic categories and P is supervised by the ground truth during training process. Then we apply a 1×1 convolution on P to get the spatial mask.

$$P_{nij} = \sum_{m=1}^{C} \omega'_{nm} \cdot F_{mij}, \qquad n = 1, ..., c_0, \tag{5}$$

$$g_{ij} = \sigma(\sum_{n=1}^{c_0} \theta_n \cdot P_{nij}), \tag{6}$$

Gated Scale-Transfer Operation

supervised GSTO



(c) supervised GSTO

Multi-scale Backbone with GSTO

The recently proposed multi-scale backbone HRNet [3, 2] has shown impressive results in pixel labeling tasks including semantic segmentation. With our proposed GSTO, we build an advanced backbone named GSTO-HRNet

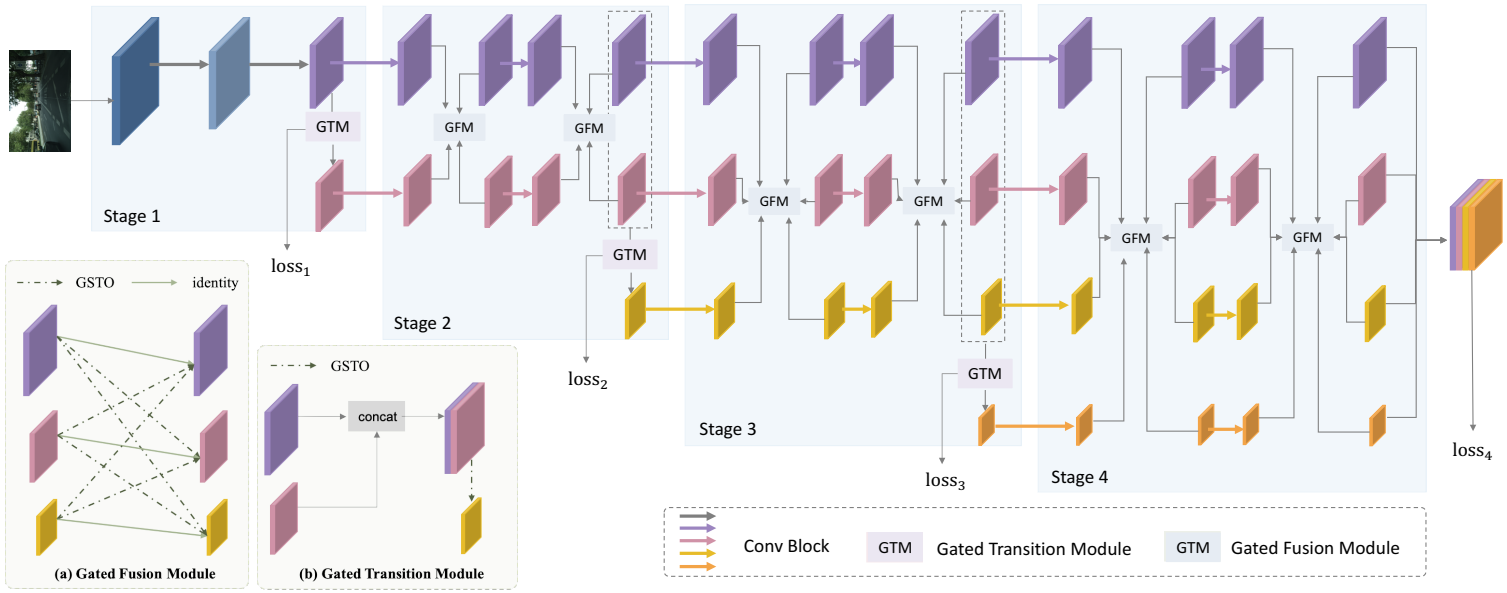


Figure: The pipeline of GSTO-HRNet, the GSTO-advanced multi-scale backbone. The GFM and GTM are GSTO-based modules for multi-scale feature fusion and generation, respectively.

Multi-scale Modules with GSTO

Traditional classification backbone can gain improvement by applying GSTO to multi-scale aggregation modules like Pyramid Pooling Module (PPM) [4] and Atrous Spatial Pyramid Pooling (ASPP) [1].

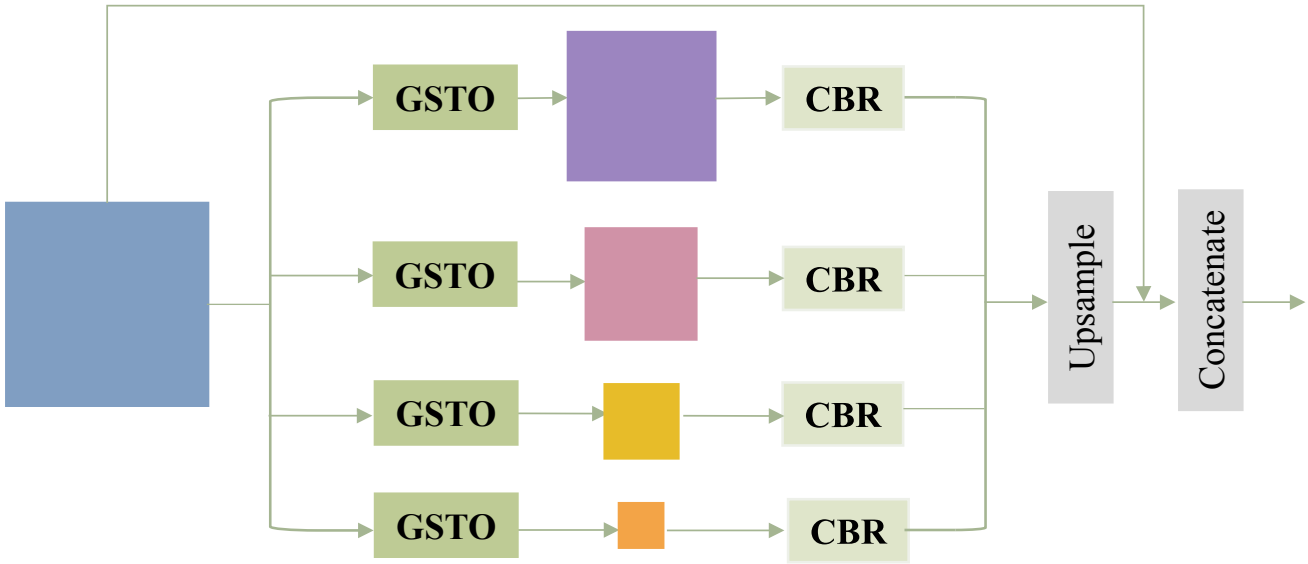


Figure: GSTO-based Pyramid Pooling Module: an example to advance multi-scale aggragation modules with the proposed GSTO. **CBR** represents Conv+BN+ReLU.

Outline

Results on GSTO-HRNet

Table: The increments of parameters and GFLOPs from HRNetV2 to our GSTO-HRNet and the mIoU comparison on Cityscapes `val.`(single scale and no flipping, not using OHEM during training).

Method	Backbone	#Param.	incre.	GFLOPs	incre.	mIoU
HRNetV2	HRNetV2-W18	3.92M	▲0.67%	71.6	▲3.9%	76.2/75.9(*impl.*)
Ours	GSTO-HRNet-W18	3.95M		74.4		**77.3**(1.1/1.4 ↑)
HRNetV2	HRNetV2-W48	65.78M	▲0.23%	696.2	▲2.6%	80.9/80.2(*impl.*)
Ours	GSTO-HRNet-W48	65.93M		714.0		**82.1**(1.2/1.9 ↑)

Results on Multi-scale Modules with GSTO

Table: Improvement on multi-scale aggregation modules.

Method	PPM	ASPP
Baseline	76.5	74.9
Baseline(w/ sup)	76.9(0.4 ↑)	75.1(0.2 ↑)
Baseline+GSTO(w/o sup)	77.3(0.8 ↑)	76.3(1.4 ↑)
Baseline+GSTO(w/ sup)	77.8(1.3 ↑)	76.9(2.0 ↑)

Cityscapes

Table: Comparison with state-of-the-art segmentation results on Cityscapes `test`.

Method	Backbone	mIoU	iIoU cla.	IoU cat.	iIoU cat.
Model learned on the `train` set					
PSPNet	Dilated-ResNet-101	78.4	56.7	90.6	78.6
PSANet	Dilated-ResNet-101	78.6	-	-	-
AAF	Dilated-ResNet-101	79.1	-	-	-
HRNetV2	HRNetV2-W48	80.4	59.2	91.5	80.8
ACFNet	ResNet-101	80.8	-	-	-
Our approach	GSTO-HRNet-W48	**81.8**	**62.3**	**92.1**	**81.7**
Model learned on the `train+valid` set					
DeepLab	Dilated-ResNet-101	70.4	42.6	86.4	67.7
RefineNet	ResNet-101	73.6	47.2	87.9	70.6
DFN	ResNet-101	79.3	-	-	-
PSANet	Dilated-ResNet-101	80.1	-	-	-
DenseASPP	WDenseNet-161	80.6	59.1	90.9	78.1
SPGNet	2×ResNet-50	81.1	-	-	-
HRNetV2	HRNetV2-W48	81.6	61.8	92.1	82.2
ACFNet	ResNet-101	81.8	-	-	-
Our approach	GSTO-HRNet-W48	**82.4**	**63.8**	**92.4**	**83.3**

LIP

Table: Semantic segmentation results on LIP. *N* denotes not using any extra information, e.g., pose or edge.

Method	Backbone	Extra.	Pixel acc.	Avg. acc.	mIoU
Attention+SSL	VGG16	Pose	84.36	54.94	44.73
DeepLabV3+	Dilated-ResNet-101	-	84.09	55.62	44.80
MMAN	Dilated-ResNet-101	-	-	-	46.81
SS-NAN	ResNet-101	Pose	87.59	56.03	47.92
MuLA	Hourglass	Pose	88.50	60.50	49.30
JPPNet	Dilated-ResNet-101	Pose	86.39	62.32	51.37
CE2P	Dilated-ResNet-101	Edge	87.37	63.20	53.10
HRNetV2	HRNetV2-W48	N	88.21	67.43	55.90
Our approach	GSTO-HRNet-W48	N	**88.38**	**68.36**	**57.37**

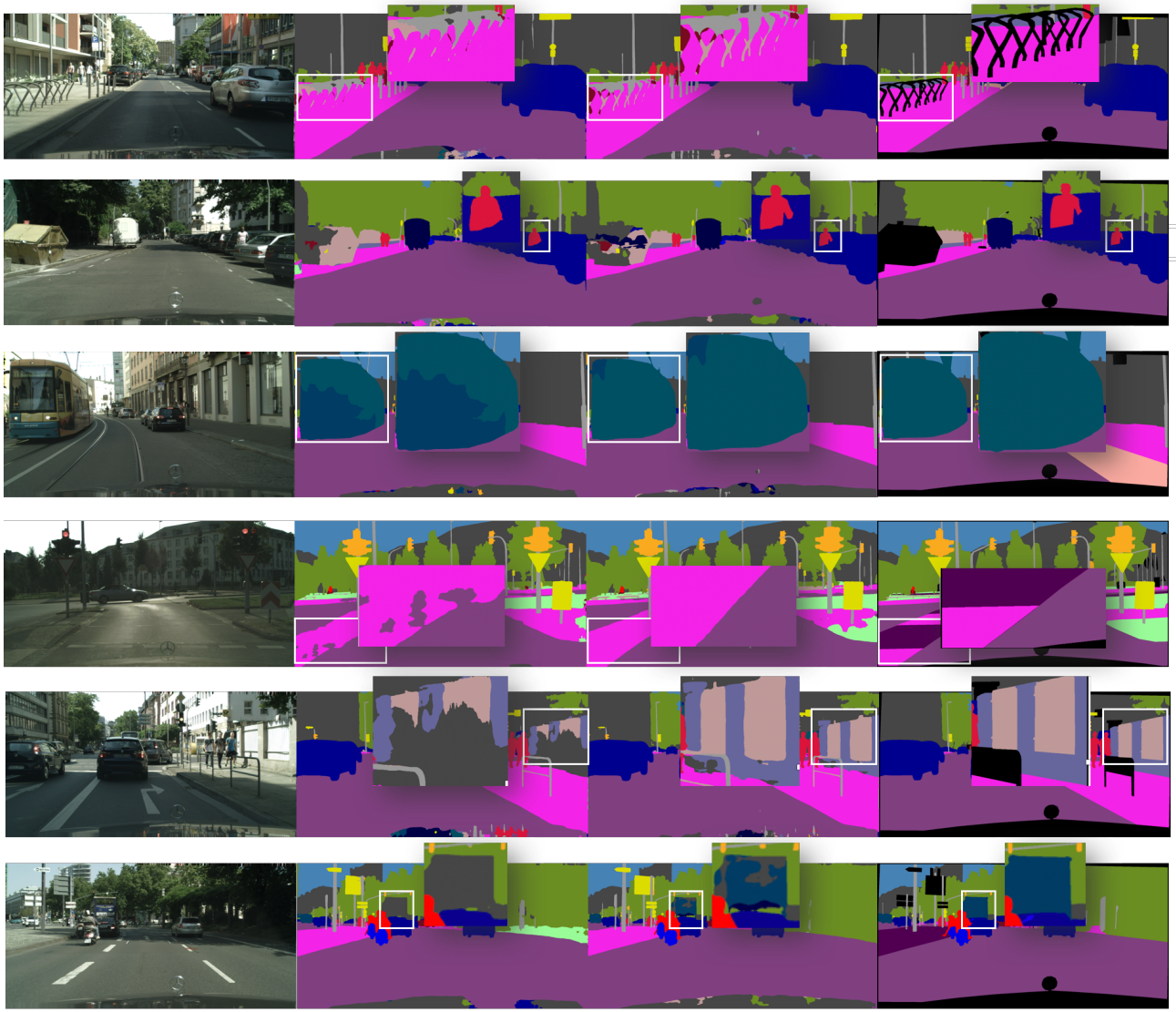
Pascal Context

Table: Semantic segmentation results on PASCAL-context. The methods are evaluated on 59 classes and 60 classes.

Method	Backbone	mIoU (59 classes)	mIoU (60 classes)
FCN-8s	VGG-16	-	35.1
BoxSup	-	-	40.5
DeepLab-v2	Dilated-ResNet-101	-	45.7
RefineNet	ResNet-152	-	47.3
PSPNet	Dilated-ResNet-101	47.8	-
Ding et al.	ResNet-101	51.6	-
EncNet	Dilated-ResNet-101	52.6	-
HRNetV2	HRNetV2-W48	54.0	48.3
Our approach	GSTO-HRNet-W48	**54.3**	**48.5**

Outline

Visualization



References

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille.

Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs.

PAMI, 2018.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang.

Deep high-resolution representation learning for human pose estimation.

2019.

Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang.

High-resolution representations for labeling pixels and regions.

arXiv:1904.04514, 2019.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia.

Pyramid scene parsing network.