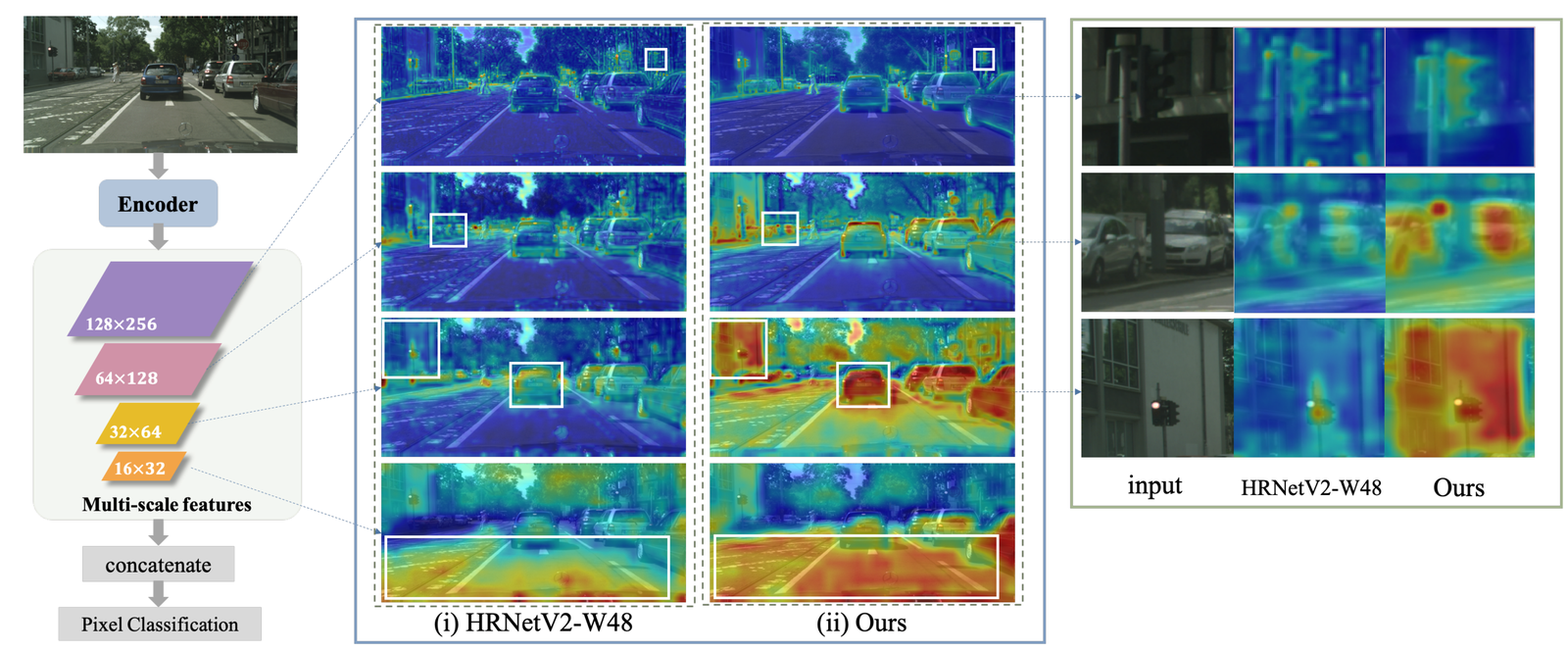
GSTO: Gated Scale-Transfer Operation for Multi-Scale Feature Learning in Semantic Segmentation

Zhuoying Wang, Yongtao Wang*, Zhi Tang, Yangyan, Ying Chen, Haibin Ling and Weisi Lin
Peking University, Alibaba Group, Brook University, Nanyang Technological University
Email: {wzypku, wyt, tangzhi}@pku.edu.cn, {yangyan.lyy, chenying.ailab}@alibaba-inc.com, hling@cs.stonybrook.edu, wslin@ntu.edu.sg

1. Introduction

Existing CNN-based methods for semantic segmentation heavily depend on multi-scale features to meet the requirements of both semantic comprehension and detail preservation. State-of-the-art segmentation networks widely exploit conventional scale-transfer operations, i.e., up-sampling and down-sampling to learn multi-scale features. In this work, we find that these operations lead to scale-confused features and suboptimal performance because they are spatial-invariant and directly transit all feature information cross scales without spatial selection. To address this issue, we propose the Gated Scale-Transfer Operation (GSTO) to properly transit spatial-filtered features to another scale. Specifically, GSTO can work either with or without extra supervision. Unsupervised GSTO is learned from the feature itself while the supervised one is guided by the supervised probability matrix. Both forms of GSTO are lightweight and plug-and-play, which can be flexibly integrated into networks or modules for learning better multi-scale features. In particular, by plugging GSTO into HRNet, we get a more powerful backbone (namely GSTO-HRNet) for pixel labeling, and it achieves new state-of-the-art results on multiple benchmarks for semantic segmentation including Cityscapes, LIP, and Pascal Context, with a negligible extra computational cost. Moreover, experiment results demonstrate that GSTO can also significantly boost the performance of multi-scale feature aggregation modules like PPM and ASPP.

2. Scale Confusion



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

Figure 1: 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. The comparison demonstrates that our approach obtains more discriminate and scale-aware features, where small objects like "traffic light" and object boundaries are more precisely highlighted in the high-resolution feature map, while medium-size objects like "car" and far-away "building" as well as large objects like "road" and nearby "car" are better focused in low-resolution feature maps. On the contrast, HRNetV2 suffers from feature-confusion, that is, some parts of large objects incorrectly fire high activation responses on the high-resolution features and large objects are insufficiently focused on the low-resolution features.

3. GSTO

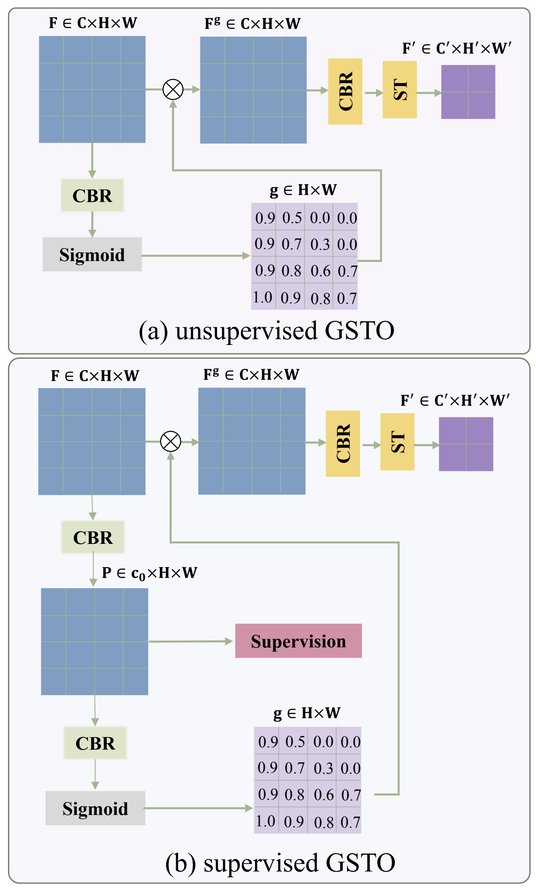


Figure 2: Structures of the proposed unsupervised GSTO and supervised GSTO. **CBR** represents Conv+BN+ReLU, used to change the channel size, if needed, and **ST** refers to conventional scale-transfer operation including down-sampling and up-sampling.

5. Experiment

Datasets Cityscapes is a large-scale dataset focusing on semantic understanding of urban street scenes, containing $5,000$ pixel-level annotated scene images divided into $2,975/500/1,525$ images for training, validation, and testing, respectively. For pixel-level labeling, there are 30 classes annotated, and 19 of them used for evaluation.

Table 1: 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	-	-	-
PAN	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					
GridNet	-	69.5	44.1	87.9	71.1
DeepLab	Dilated-ResNet-101	70.4	42.6	86.4	67.7
FRRN	-	71.8	45.5	88.9	75.1
DepthSeg	Dilated-ResNet-101	78.2	-	-	-
RefineNet	ResNet-101	73.6	47.2	87.9	70.6
BiSeNet	ResNet-101	78.9	-	-	-
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**

4. Method

In the proposed GSTOs (Figure 2(b) and (c)), a spatially gated feature F^g is produced first.

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

unsupervised GSTO As Figure 2(b), the element of the gate g_{ij} is calculated from the original feature F:

$$g_{ij} = \sigma(\sum_{m=1}^{C} \rho_m \cdot F_{mij}), \tag{2}$$

supervised GSTO As Figure 2(c), a lightweight 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.

$$P_{nij} = \sum_{m=1}^{C} \omega'_{nm} \cdot F_{mij}, \qquad n = 1, ..., c_0, \tag{3}$$

$$g_{ij} = \sigma(\sum_{n=1}^{c_0} \theta_n \cdot P_{nij}), \tag{4}$$