— Supplementary material —
Semi-Supervised Semantic Segmentation with Pixel-Level Contrastive Learning from a Class-wise Memory Bank

Inigo Alonso[1] Alberto Sabater[1] David Ferstl[2] Luis Montesano[1,3] Ana C. Murillo[1]

[1]RoPeRT group, at DIIS - I3A, Universidad de Zaragoza, Spain

[2]Magic Leap, Zürich, Switzerland

[3]Bitbrain, Zaragoza, Spain

{inigo, asabater, montesano, acm}@unizar.es, dferstl@magicleap.com

This document provides additional details of our work. It provides qualitative results of segmentation results obtained with our method are shown. Besides, the per-class evaluation of the benchmarking is provided, which expands the mIoU evaluation from the main paper.

1. Qualitative Results

This section provides additional qualitative results for our semi-supervised semantic segmentation method on the Cityscapes and Pascal-VOC datasets.

To further analyze the performance of our method in a more visual manner, and to make it easier for future visual comparison with future work, we show 20 segmentation results for both the Cityscapes and Pascal-VOC datasets. We visually compare the result of our method with the reference fully supervised set-up and the ground truth. Our method uses the most challenging labeled ratio: $\frac{1}{30}$ for the Cityscapes and $\frac{1}{50}$ for Pascal-VOC.

Figures 1 and 2 show the visual results for the Cityscapes. We can appreciate that for several images, the results are very similar to the segmentation achieved by the fully supervised version. Even for difficult classes (traffic light, traffic sign, pole, person, rider) our method trained on only $\frac{1}{30}$ labels shows really good performance. Usually the fully supervised set-up shows better performance on borders and small shape details, which is probably the remaining challenge for semi-supervised approaches. Note that black pixels from the labels are ignored.

Figures 3, 4 and 5 show the visual results for the Pascal-VOC. This dataset contain more simple scenes compared to Cityscapes samples (urban scenarios) where usually only one or two objects are segmented. Similarly to the Cityscapes comparison, we can see that our method trained on only $\frac{1}{50}$ shows similar performance to the fully supervised setting. Most cases where the fully supervised model outperforms our method are again in borders and details. In this data we can also appreciate advantage of the fully supervised approach for difficult classes like chair or potted plant. Note that white pixels from the labels are ignored.

2. Per-class Benchmark Evaluation

In this section, we show the benchmarking evaluation detailing the per-class performance. Previous work, in particular the approaches considered reference baselines for our work, do not specify this per-class evaluation, only reporting the mean Intersection over Union (mIoU). Therefore, no comparison can be made with other works. However, this allows future works to be able to compare in a per-class level.

As specified in the main paper, the results are the mean of three executions on different training data splits. Table 1 shows the per-class performance for semi-supervised semantic segmentation on the Cityscapes benchmark and Table 2 on the Pascal-VOC benchmark. Table 3 details the per-class evaluation for the semi-supervised domain adaptation for semantic segmentation. This table is evaluated on the Cityscapes and it is trained on the full labeled GTA5 dataset and on the partially labeled Cityscapes dataset.

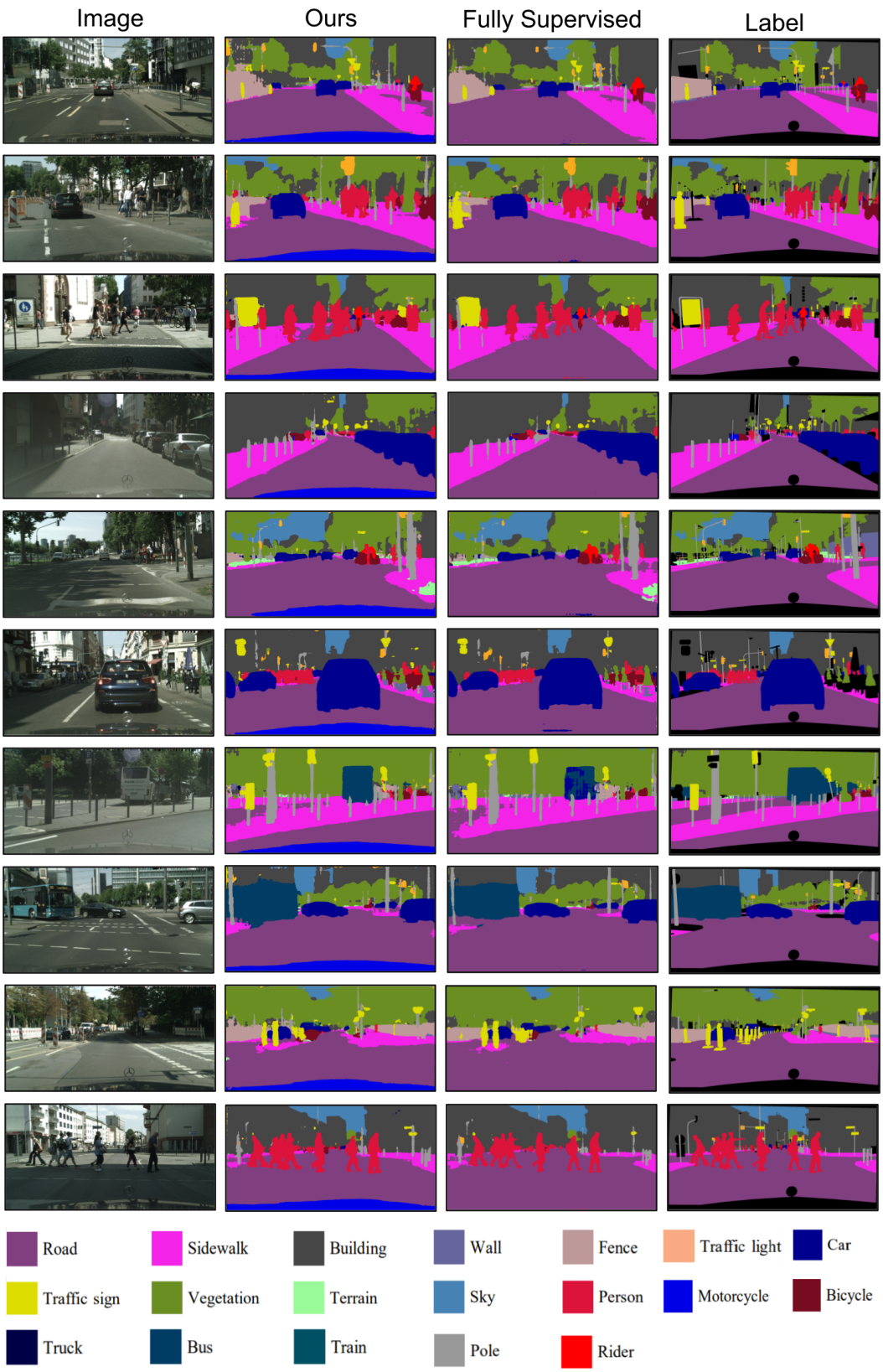

Figure 1. Qualitative results on Cityscapes. Models are trained using Deeplabv2 with ResNet-101. From left to right: Image, our method (trained on $\frac{1}{30}$ labels), fully supervised training and, the corresponding labels.

Figure 2. Qualitative results on Cityscapes. Models are trained using Deeplabv2 with ResNet-101. From left to right: Image, our method (trained on $\frac{1}{30}$ labels), fully supervised training and, the corresponding labels.

Figure 3. Qualitative results on Pascal-VOC. Models are trained using Deeplabv2 with ResNet-101. From left to right: Image, our method (trained on $\frac{1}{50}$ labels), fully supervised training and, the corresponding labels.

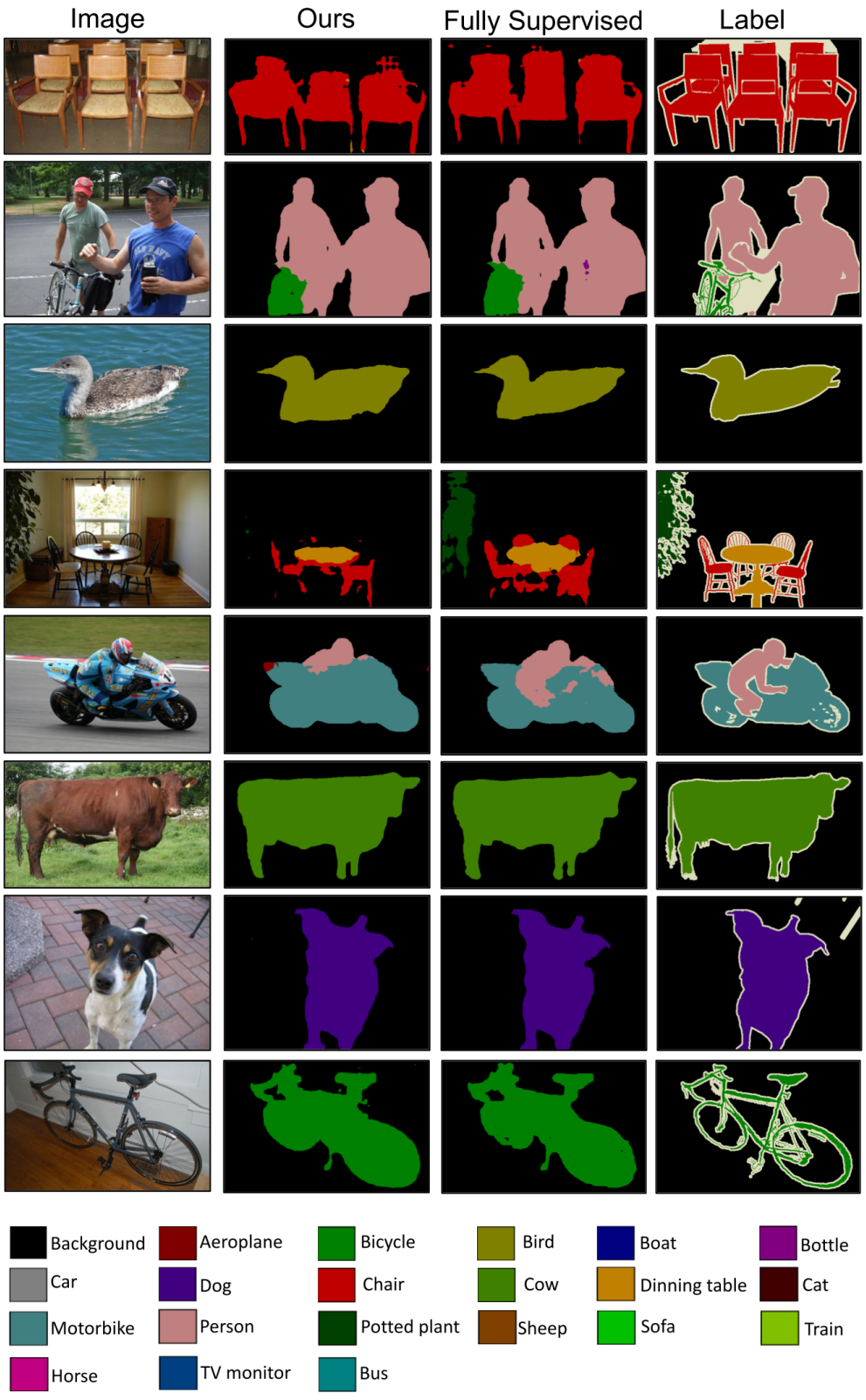

Figure 4. Qualitative results on Pascal-VOC. Models are trained using Deeplabv2 with ResNet-101. From left to right: Image, our method (trained on $\frac{1}{50}$ labels), fully supervised training and, the corresponding labels.

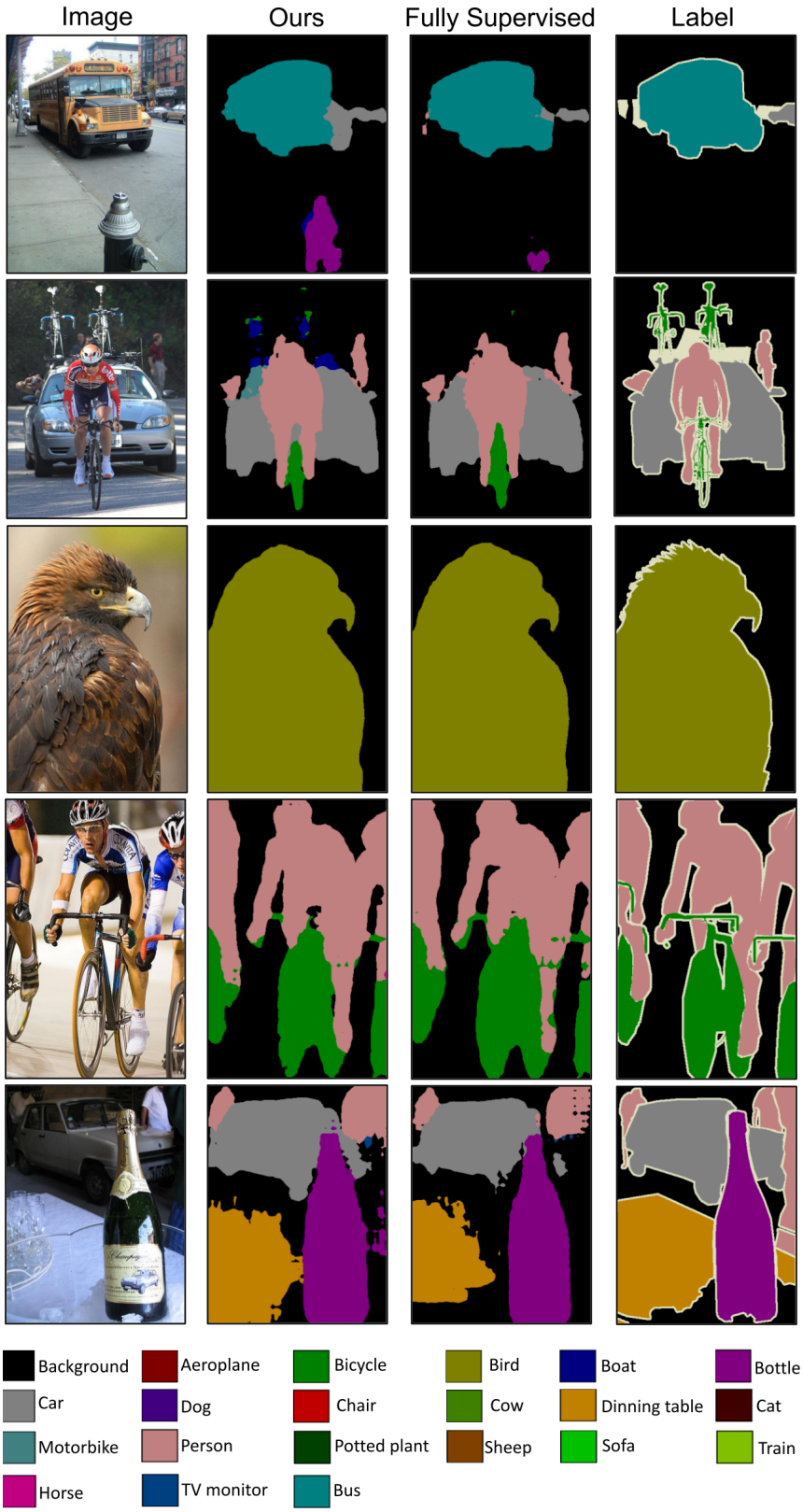

Figure 5. Qualitative results on Pascal-VOC. Models are trained using Deeplabv2 with ResNet-101. From left to right: Image, our method (trained on $\frac{1}{50}$ labels), fully supervised training and, the corresponding labels.

Table 1. Per-class performance (IoU) for semi-supervised semantic segmentation. Train on Cityscapes *train* split and evaluated on Cityscapes *val* set. Different configurations of our method are compared.

Label Ratio	**Pre-training**	**Architecture**	**mIoU**	Road IoU	Sidewalk IoU	Building IoU	Wall IoU	Fence IoU	Pole IoU	Traffic light IoU	Traffic sign IoU	Vegetation IoU	Terrain IoU	Sky IoU	Person IoU	Rider IoU	Car IoU	Truck IoU	Bus IoU	Train IoU	Motorcycle IoU	Bicycle IoU
$\frac{1}{30}$	I	Dv2	58.0	95.9	70.2	85.1	37.6	36.8	37.0	43.3	52.9	86.6	44.6	90.2	62.2	40.0	87.7	44.8	60.7	26.8	37.5	63.0
$\frac{1}{30}$	C	Dv2	59.4	96.0	71.5	86.8	42.3	36.3	41.0	46.2	55.3	86.5	44.8	90.6	65.2	41.7	88.5	46.1	55.0	31.9	40.7	63.1
$\frac{1}{30}$	I	Dv3+	64.8	96.9	77.8	90.2	40.4	41.6	55.6	56.3	67.5	90.0	50.3	90.8	74.2	53.4	91.1	45.6	65.3	28.5	50.0	68.7
$\frac{1}{15}$	I	Dv2	59.9	96.4	71.8	85.9	41.0	38.8	38.4	43.5	53.4	86.7	47.3	90.3	63.5	42.9	88.6	48.5	63.3	62.8	41.0	63.7
$\frac{1}{8}$	I	Dv2	63.0	97.0	73.5	87.6	48.2	41.9	40.1	44.3	55.8	87.7	51.3	90.9	65.6	44.5	89.4	58.0	68.4	42.7	45.8	63.6
$\frac{1}{8}$	C	Dv2	64.4	97.1	74.6	88.1	48.3	39.5	44.0	49.3	60.4	88.7	53.3	92.0	67.5	47.8	90.8	67.7	70.6	33.1	46.6	64.2
$\frac{1}{8}$	I	Dv3+	70.0	97.4	79.9	90.2	42.7	45.0	56.6	58.2	70.3	91.1	59.3	92.2	76.0	54.8	93.1	65.5	74.8	56.6	56.8	70.4
$\frac{1}{6}$	I	Dv2	63.7	87.2	74.2	87.7	49.9	43.5	40.3	44.1	56.8	88.0	52.1	90.9	67.2	45.3	89.8	63.2	69.4	45.1	46.3	62.7
$\frac{1}{4}$	I	Dv2	64.8	97.4	75.3	87.6	50.1	46.0	40.8	44.4	57.6	88.2	53.1	91.0	67.5	46.7	90.2	67.5	71.3	51.7	42.3	61.6
$\frac{1}{4}$	C	Dv2	65.9	97.2	76.1	87.9	47.8	46.1	44.8	48.6	61.0	89.1	54.9	91.8	68.5	49.6	90.9	69.0	73.6	48.2	43.5	63.3
$\frac{1}{4}$	I	Dv3+	71.6	97.5	80.4	90.6	45.6	47.3	57.2	56.8	69.4	90.9	58.7	91.7	76.1	52.0	93.1	64.6	76.2	60.5	55.7	68.8
$\frac{1}{3}$	I	Dv2	65.1	97.4	75.7	87.7	50.8	45.6	41.1	45.8	58.4	88.6	53.0	91.1	67.8	47.3	90.2	68.1	72.5	50.9	44.2	61.4

I: ImageNet, **C**: COCO, **Dv2**: DeeplabV2 with ResNet-101 backbone, **Dv3+**: DeeplabV3+ with ResNet-50 backbone

Table 2. Per-class performance (IoU) for semi-supervised semantic segmentation. Train on Pascal-VOC *train* split and evaluated on Pascal-VOC *val* set. Different configurations of our method are compared.

Label Ratio	**Pre-training**	**Architecture**	**mIoU**	Background IoU	Aeroplane IoU	Bicycle IoU	Bird IoU	Boat IoU	Bottle IoU	Bus IoU	Car IoU	Cat IoU	Chair IoU	Cow IoU	Dining Table IoU	Dog IoU	Horse IoU	Motorbike IoU	Person IoU	Potted plant IoU	Sheep IoU	Sofa IoU	Train IoU	TV monitor IoU
$\frac{1}{50}$	I	Dv2	65.4	91.5	75.4	37.8	82.5	43.1	65.9	88.5	80.3	83.4	22.9	75.9	41.2	69.3	70.1	72.0	77.9	48.6	66.8	44.3	76.2	60.8
$\frac{1}{50}$	C	Dv2	67.9	92.1	77.3	35.5	83.2	58.2	66.8	89.6	79.9	83.3	25.3	76.0	45.4	77.3	76.6	72.8	79.8	49.2	73.1	48.3	77.8	59.4
$\frac{1}{50}$	I	Dv3+	63.4	91.3	74.9	39.8	84.6	39.1	67.6	88.7	82.3	86.2	29.0	74.7	37.8	64.9	66.1	72.8	79.8	50.1	61.4	47.2	78.7	63.9
$\frac{1}{20}$	I	Dv2	67.8	91.9	76.5	37.5	83.4	54.0	67.2	89.3	82.5	84.9	26.7	76.7	43.6	75.5	75.1	73.9	79.5	49.6	71.8	44.7	77.6	61.3
$\frac{1}{20}$	C	Dv2	70.0	93.1	81.4	39.4	85.3	55.1	69.9	89.3	84.7	88.0	30.4	79.9	45.7	78.9	78.6	76.3	81.0	51.5	76.2	46.1	78.0	62.1
$\frac{1}{20}$	I	Dv3+	69.1	92.2	81.4	42.5	81.8	60.4	64.8	88.6	81.5	84.2	22.7	78.8	52.1	79.7	78.6	78.2	78.7	52.0	73.4	40.5	75.9	62.2
$\frac{1}{8}$	I	Dv2	69.9	92.8	81.7	39.2	84.7	57.0	69.1	89.0	82.8	86.4	28.2	80.3	48.0	78.9	79.3	77.3	80.8	52.7	74.1	44.4	78.9	61.3
$\frac{1}{8}$	C	Dv2	71.6	93.4	82.5	38.6	86.9	63.1	74.0	82.3	84.8	86.1	31.0	81.2	48.5	80.7	79.7	77.1	83.3	53.8	75.2	45.8	81.1	53.7
$\frac{1}{8}$	I	Dv3+	71.8	93.4	84.2	39.5	85.9	61.4	67.6	91.3	82.2	87.0	30.1	82.6	53.3	81.4	81.1	78.5	80.5	57.9	76.7	46.1	82.6	64.3

I: ImageNet, **C**: COCO, **Dv2**: DeeplabV2 with ResNet-101 backbone, **Dv3+**: DeeplabV3+ with ResNet-50 backbone

Table 3. Per-class performance (IoU) for semi-supervised domain adaptation semantic segmentation. Train on the GTA5 dataset as the full labeled dataset and Cityscapes as the partially labeled dataset. Evaluated on Cityscapes *val* set. Different configurations of our method are compared. All configurations use the Deeplabv3+ with ResNet-50 backbone and are ImageNet pre-trained.

Label Ratio	**mIoU**	Road IoU	Sidewalk IoU	Building IoU	Wall IoU	Fence IoU	Pole IoU	Traffic light IoU	Traffic sign IoU	Vegetation IoU	Terrain IoU	Sky IoU	Person IoU	Rider IoU	Car IoU	Truck IoU	Bus IoU	Train IoU	Motorcycle IoU	Bicycle IoU
$\frac{1}{30}$	59.9	95.4	70.3	86.5	43.9	41.4	37.1	41.8	54.5	86.6	46.1	89.5	62.9	42.7	88.0	51.2	61.8	36.4	39.8	62.6
$\frac{1}{15}$	62.0	96.3	71.4	86.8	47.6	39.8	38.0	42.2	56.5	87.1	50.4	90.0	65.3	43.6	89.8	63.9	65.2	40.7	42.4	61.5
$\frac{1}{6}$	64.2	96.4	72.5	87.0	48.8	44.1	38.4	42.5	57.6	87.1	52.5	90.1	66.0	46.9	90.2	69.7	70.3	51.3	45.9	61.8
$\frac{1}{3}$	65.6	96.4	72.3	87.5	50.4	47.5	40.8	43.7	59.0	87.6	52.4	791.0	66.1	46.4	90.3	72.9	74.2	57.9	47.5	62.2