Improving Semantic Segmentation via Self-Training

Yi Zhu, Zhongyue Zhang, Chongruo Wu, Zhi Zhang, Tong He, Hang Zhang,
R. Manmatha, Mu Li, Alexander Smola

Contribution

- Introduces new, semi-supervised self-training method for semantic segmentation
  - Self-training method is similar to *pseudo-labelling* in image classification
  - Authors show efficacy of in cross-domain generalization
  - Also propose and test faster training schedule for self-training

Motivation

- Semantic Image Segmentation (classifying each pixel in an image) is an important study in deep learning/computer vision.

- Fully supervised training for semantic segmentation is often impractical, as it requires pixel-by-pixel labelling for thousands of images.

- Effective, generalizable, unsupervised methods for segmentation are desired.

Motivation Cont.

- **Key idea in this paper includes adapting pseudo-label type training to segmentation on driving scene images.**

- **Pseudo-Labeling**: a semi-supervised training method
  - Train the network with labeled data first
  - Make network predictions on unlabeled data
  - Assign predicted labels to samples with high-confidence predictions
  - Retrain with labeled data

Teacher/Student Paradigm w/ Pseudo-Labeling

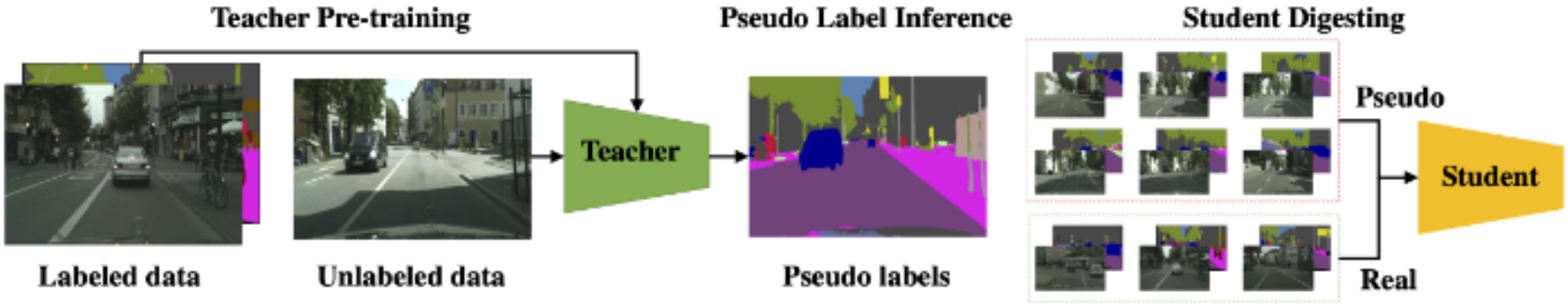

Fig. 1. Overview of our self-training framework. (a) Train a teacher model on labeled data. (b) Generate pseudo labels on a large set of unlabeled data. (c) Train a student model using the combination of both real and pseudo labels

Addressing Noise in Pseudo-Labeling

- Pseudo-labeling often makes little difference in accuracy, because synthesized pseudo-labels are too noisy.
- Part of this noise is due to *class imbalance*.
  - Teacher model will bias toward already overrepresented classes, amplifying imbalance.
  - Ex: teacher trained on 360:1 road/motorcycle examples. Pseudo-labels have >1000:1 road/motorcycle ratio.

Solution: Centroid Sampling

- Augment data by cropping around the centroid of a certain class.
  - This ensures that all instances of a class can be seen in a training epoch.
- Centroid sampling is done with all classes, so that the real-world distribution is approximately the same.

Addressing Long Training Times

- Training with many pseudo-labels and large crops (800x800) can take weeks.
- Reducing crops size significantly hurts results.
  - Loss of global context

Solution: Changing image size during training schedule

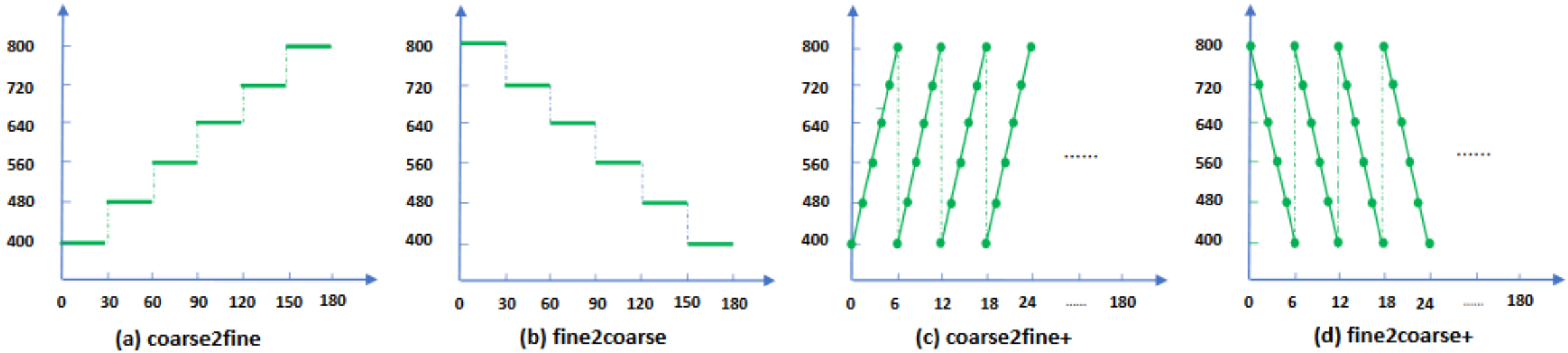

Fig. 2. Overview of our proposed fast training schedules. x-axis is the epoch number and y-axis is the crop size. See texts in Section 3.3 for more details

Experimental Objectives

- Demonstrate performance improvement from pseudo-labeling with centroid sampling self-training in semantic segmentation.
- Demonstrate efficacy of accelerated training schedule.
- Demonstrate application of self-training method to cross-domain generalization.
- For most experiments: DeepLabV3_ResNeXt50 backbone for network architecture

Marginal Performance Increase with Self-Training

	Fine	Coarse*	Mapillary*	mIoU (%)
Teacher	✓			78.1
Student	✓	✓		79.0
Student	✓	✓	✓	**79.3**

Table 1. Performance of our baseline and the results of self-training. Fine: using Cityscapes fine annotations. Coarse* and Mapillary* means we only use the images from Cityscapes coarse dataset and Mapillary dataset, without using their labels

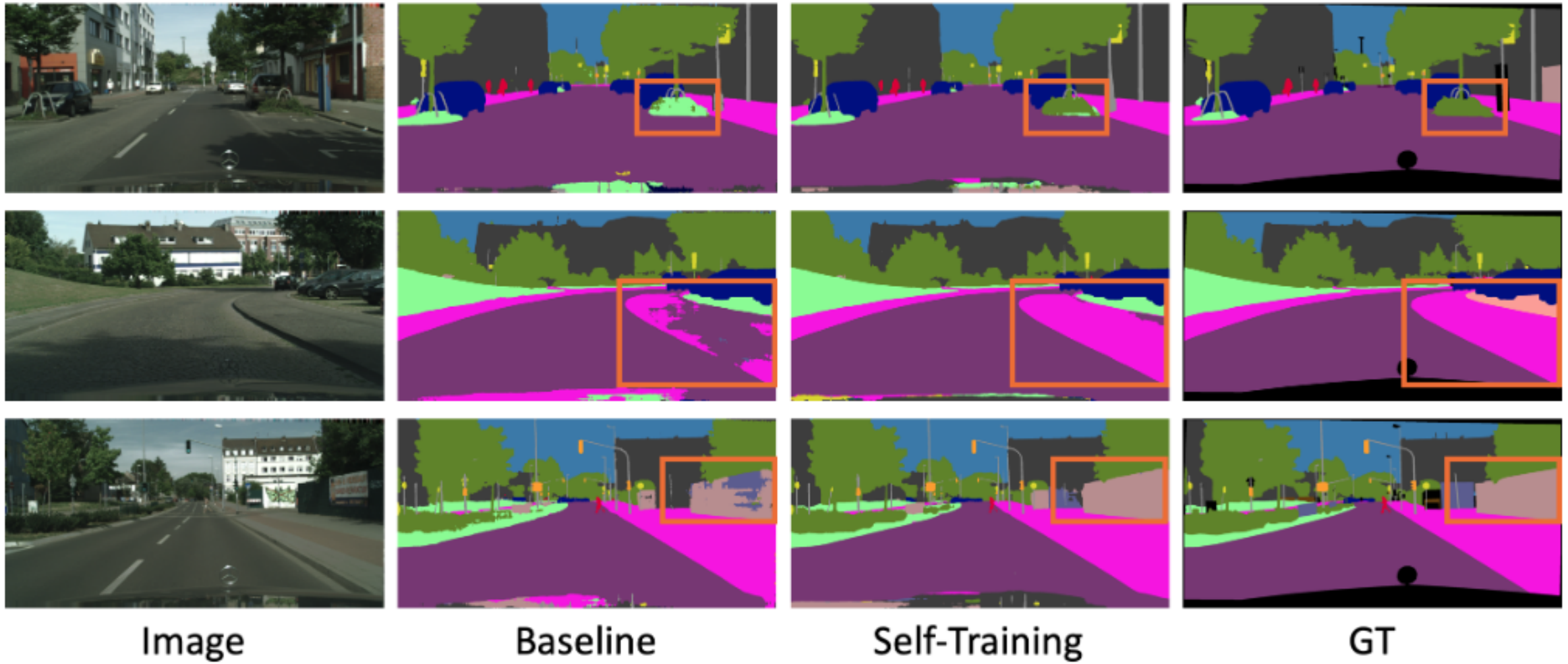

Fig. 4. Visual comparisons on Cityscapes. We demonstrate that self-training can effectively handle class confusion, such as between tree and vegetation (row 1), road and sidewalk (row 2), wall and fence (row 3).

Increasing pseudo-labels improves performance

(a) Self-training with pseudo labels

	Real	Pseudo	w/o CS	w CS
Teacher	3K	-	78.1	78.1
Student	1.5K	1.5K	78.4	79.3
Student	1.5K	4.5K	78.7	79.7
Student	1.5K	7.5K	78.9	79.9
Student	1.5K	10.5K	78.7	**80.0**
VPLR [79]	43K	-	-	79.8

(b) Duplicate real labels

	Real	mIoU (%)
Teacher	3K	78.1
Teacher	6K	78.9
Teacher	9K	**79.1**
Teacher	12K	79.0

Table 2. Ablation study on the ratio between pseudo and real labels. We find that increasing the ratio of pseudo labels improves the segmentation accuracy, and even outperforms a model pre-trained on 43K real labels [79]. CS: centroid sampling

Self-training can improve results in various student models

	Network	Backbone	# Real	mIoU (%)
Teacher	DeepLabv3+	ResNeXt50	3K	78.1
Baseline [79]	DeepLabV3+	WideResNet38	3K	80.5
Mapillary Pre-trained [79]	DeepLabV3+	WideResNet38	43K	81.5
Self-training(ours)	DeepLabV3+	WideResNet38	3K	**82.2**
Baseline [49]	FastSCNN	-	3K	68.6
Mapillary Pre-trained [49]	FastSCNN	-	43K	71.7
Self-training(ours)	FastSCNN	-	3K	**72.5**
Baseline [76]	PSPNet	ResNet101	3K	77.9
Mapillary Pre-trained [76]	PSPNet	ResNet101	43K	79.2
Self-training(ours)	PSPNet	ResNet101	3K	**79.9**

Table 3. Our self-training method can improve the student models irrespective of backbones and network architectures. We again outperform models pre-trained on Mapillary labeled data for all three students. # Real: the number of real labels used in training

Accelerating Training Schedule

- coarse2fine+ training schedule achieved 1.7x training speed with no accuracy loss.
- Accelerated schedule benefitted from good initialization via 'warm-up'
  - Initial 20 epochs with largest (800) crop before schedule.

	w/o warm-up	w warm-up
Baseline	80.0	80.0
coarse2fine	79.1	79.6
fine2coarse	79.4	79.8
coarse2fine+	79.5	**80.0**
fine2coarse+	79.4	79.9
speed up	1.8x	1.7x

Table 4. Comparison of fast training schedules. coarse2fine+ using crop size warm-up is able to achieve 1.7x speed up without performance degradation

Comparison to State-of-the-Art

Method	road	swalk	build.	wall	fence	pole	tlight	tsign	veg.	terrain	sky	person	rider	car	truck	bus	train	mcycle	bicycle	mIoU
DepthSeg [27]	98.5	85.4	92.5	54.4	60.9	60.2	72.3	76.8	93.1	71.6	94.8	85.2	68.9	95.7	70.1	86.5	75.5	68.3	75.5	78.2
PSPNet [76]	98.6	86.2	92.9	50.8	58.8	64.0	75.6	79.0	93.4	72.3	95.4	86.5	71.3	95.9	68.2	79.5	73.8	69.5	77.2	78.4
AAF [26]	98.5	85.6	93.0	53.8	58.9	65.9	75.0	78.4	93.7	72.4	95.6	86.4	70.5	95.9	73.9	82.7	76.9	68.7	76.4	79.1
PanoDeepLab [12]	98.7	87.2	93.6	57.7	60.8	70.8	78.0	81.2	93.8	74.1	95.7	88.2	76.4	96.0	55.3	75.1	79.6	72.1	74.0	79.4
DenseASPP [66]	98.7	87.1	93.4	60.7	62.7	65.6	74.6	78.5	93.6	72.5	95.4	86.2	71.9	96.0	78.0	90.3	80.7	69.7	76.8	80.6
SPG [11]	98.8	87.6	93.8	56.5	61.9	71.9	80.0	82.1	94.1	73.5	96.1	88.7	74.9	96.5	67.3	84.8	81.8	71.1	79.4	81.1
BFP [14]	98.7	87.0	93.5	59.8	63.4	68.9	76.8	80.9	93.7	72.8	95.5	87.0	72.1	96.0	77.6	89.0	86.9	69.2	77.6	81.4
DANet [17]	98.6	86.1	93.5	56.1	63.3	69.7	77.3	81.3	93.9	72.9	95.7	87.3	72.9	96.2	76.8	89.4	86.5	72.2	78.2	81.5
HRNetv2 [52]	98.8	87.9	93.9	61.3	63.1	72.1	79.3	82.4	94.0	73.4	96.0	88.5	75.1	96.5	72.5	88.1	79.9	73.1	79.2	81.8
ACFNet [71]	98.7	87.1	93.9	60.2	63.9	71.1	78.6	81.5	94.0	72.9	95.9	88.1	74.1	96.5	76.6	89.3	81.5	72.1	79.2	81.8
EMANet [32]	98.7	87.3	93.8	63.4	62.3	70.0	77.9	80.7	93.9	73.6	95.7	87.8	74.5	96.2	75.5	90.2	84.5	71.5	78.7	81.9
ACNet [18]	98.7	87.1	93.9	61.6	61.8	71.4	78.7	81.7	94.0	73.3	96.0	88.5	74.9	96.5	77.1	89.0	89.2	71.4	79.0	82.3
Baseline	98.7	86.7	93.6	59.0	63.1	68.6	77.1	80.4	94.1	73.7	96.0	87.5	73.0	96.2	73.2	85.6	86.5	70.4	77.1	81.4
Mapillary-pretrained	98.8	87.6	94.1	63.8	64.7	70.4	78.1	82.1	94.2	73.5	96.1	88.3	73.7	96.3	77.2	90.9	90.4	71.9	79.0	**82.7**
Self-training	98.8	87.8	94.0	61.7	64.9	71.6	78.6	82.2	94.2	74.2	96.1	88.4	74.3	96.5	76.7	90.1	90.0	72.3	79.1	**82.7**

Table 6. Per-class comparison with top-performing methods on the test set of Cityscapes. Our method outperforms all prior literature that only uses fine labels. In terms of fair comparison, our self-trained model achieves the same segmentation accuracy as a model pre-trained using Mapillary labeled data under identical settings

Generalization/Finetuning to Other Datasets

- Take student model and finetune/test on separate image set.
- Self-training pretrain achieves best performance

Method	Pretrain	mIoU (%)
RTA [23]	ImageNet	62.5
DFANet [31]	ImageNet	64.7
BiSeNet [67]	ImageNet	68.7
PSPNet [76]	ImageNet	69.1
DenseDecoder [1]	ImageNet	70.9
SKDistill [39]	ImageNet	72.3
VideoGCRF [5]	Cityscapes	75.2
SVCNet [15]	ImageNet	75.4
VPLR [79]	Cityscapes, Mapillary	79.8
Ours	Cityscapes	**81.6**

Table 5. Results on the CamVid test set. Pre-train indicates the source dataset on which the model is trained

Generalization to Domains w/ New Classes

- Demonstrates pre-training on dataset with few semantic categories (Cityscapes, 19), transfers to use on set with many categories (Mapillary, 66).
- Self-training shows greater benefit when less fine-tune training is available.

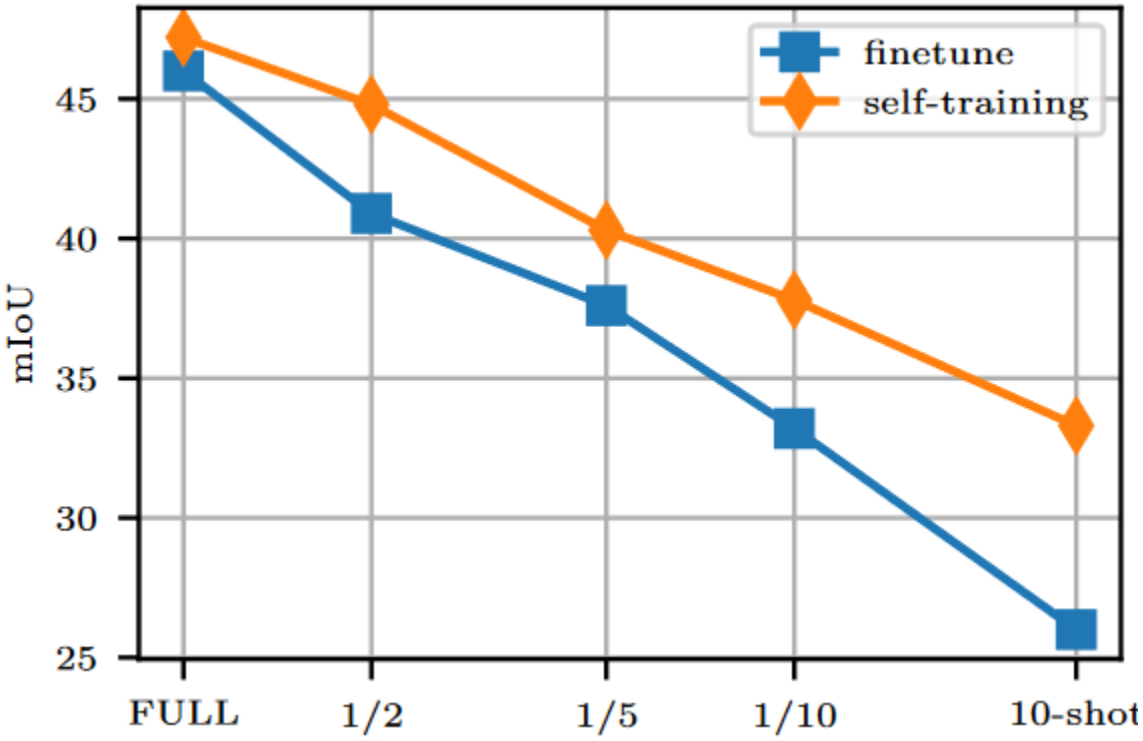

Fig. 3. Experiments on cross-domain generalization from Cityscapes to Mapillary. Full: full training set of Mapillary (18K samples). 1/2, 1/5 and 1/10 means we use 1/2, 1/5 and 1/10 of the full training set. 10-shot: 10 samples per class

Conclusion

- Work introduces new, pseudo-label based self-training method for semantic segmentation.
  - Use of centroid sampling ensures the classes are not erased in pseudo data.
  - More pseudo-labels seems to create better results
- Using an accelerated training schedule with a precise ‘warm-up’ can make training times more reasonable.
- This self-training can be used as effective pre-training for cross-domain generalizations, even with different categories.