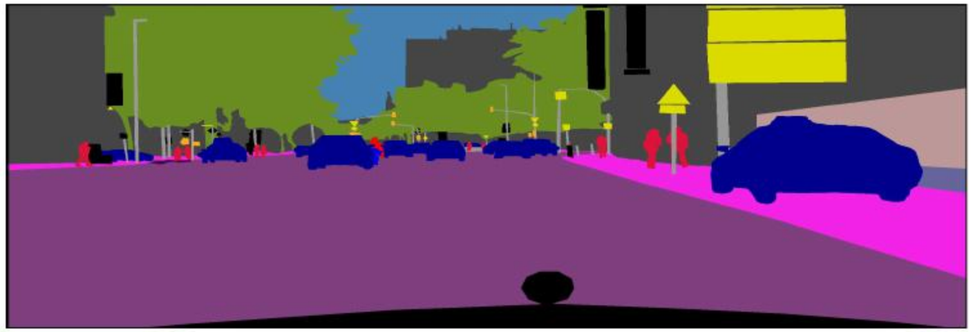
Naive-Student: Leveraging Semi-Supervised Learning in Video Sequences for Urban Scene Segmentation

Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D. Collins, Ekin D. Cubuk, Barret Zoph, Hartwig Adam, Jonathon Shlens

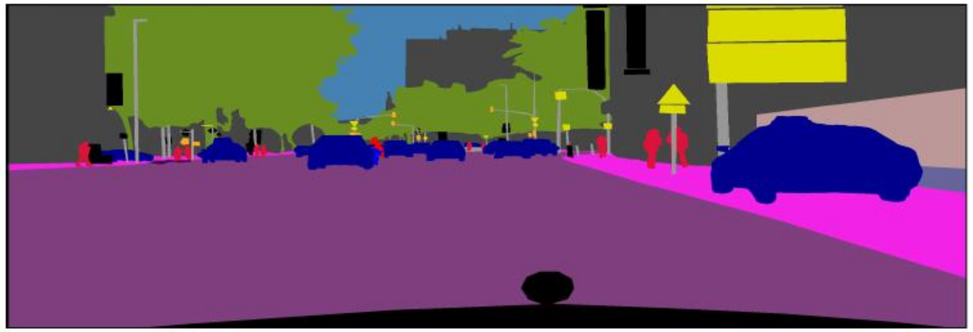
Labor-Intensive Segmentation Annotation

- Supervised-learning requires large-scale human-annotated datasets for achieving state-of-the-art results.



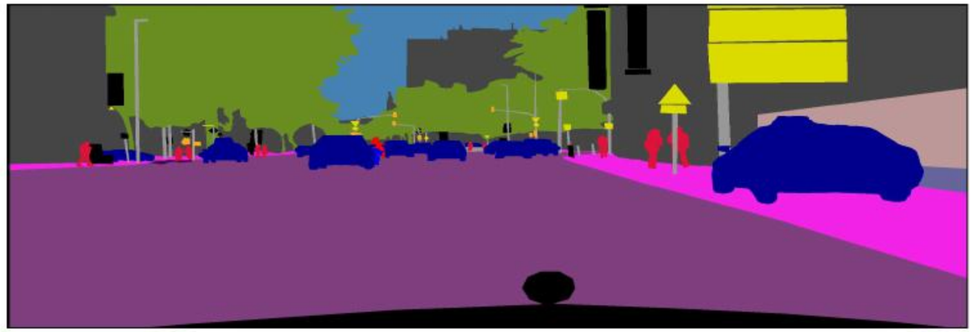
Labor-Intensive Segmentation Annotation

- Supervised-learning requires large-scale human-annotated datasets for achieving state-of-the-art results.
- In turn, the efficacy of supervised learning may be limited by the size of the human annotated dataset.



Labor-Intensive Segmentation Annotation

- Supervised-learning requires large-scale human-annotated datasets for achieving state-of-the-art results.
- In turn, the efficacy of supervised learning may be limited by the size of the human annotated dataset.
- It is extremely time-consuming to collect segmentation annotations.
  - 90 minutes per image on Cityscapes.
  - 22 hours per 1,000 segmentation masks on COCO.



- Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
- Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll´ar, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)

Limited Labeled Data + Abundant Unlabeled Data

- Small human-labeled data due to limited time and budget.
  - Cityscapes has only 3475 human-annotated images for training.

Limited Labeled Data + Abundant Unlabeled Data

- Limited human-labeled data due to limited time and budget.
  - Cityscapes has only 3475 human-annotated images for training.
- On the other hand, there are abundant unlabeled video sequences and images.
  - Cityscapes has unlabeled video sequences with > 100K frames, and extra 20K images.

Limited Labeled Data + Abundant Unlabeled Data

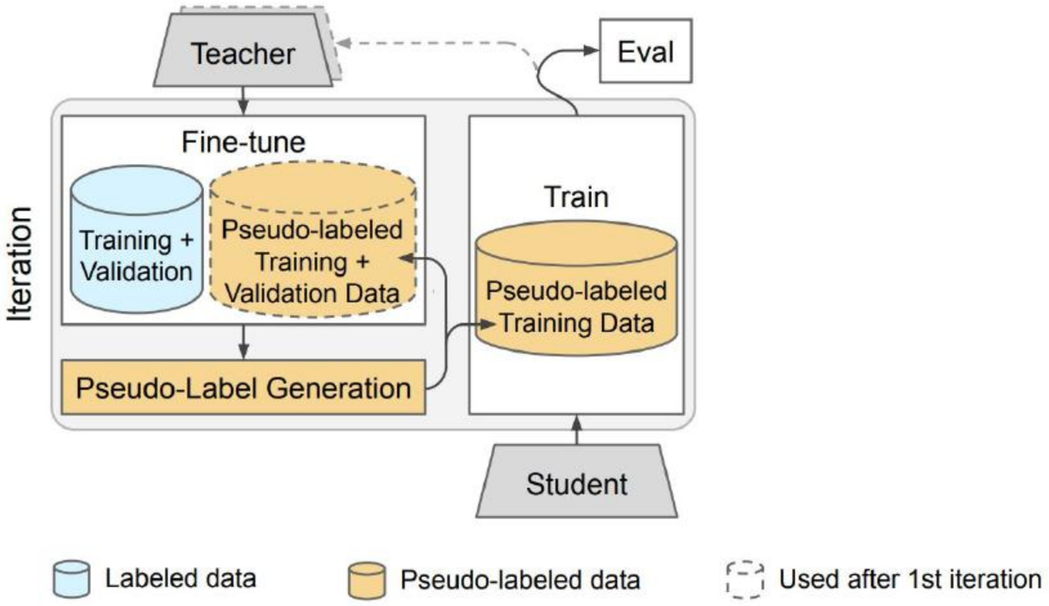
- Limited human-labeled data due to limited time and budget.
  - Cityscapes has only 3475 human-annotated images.
- On the other hand, there are abundante unlabeled video sequences and extra images.
  - Cityscapes has unlabeled video sequences with about 173.75K frames, and extra 20K images.
- Is it possible to exploit both human-labeled and unlabeled data to further improve state-of-the-art performance?

Limited Labeled Data + Abundant Unlabeled Data

- Limited human-labeled data due to limited time and budget.
  - Cityscapes has only 3475 human-annotated images.
- On the other hand, there are abundante unlabeled video sequences and extra images.
  - Cityscapes has unlabeled video sequences with about 173.75K frames, and extra 20K images.
- Is it possible to exploit both human-labeled and unlabeled data to further improve state-of-the-art performance?
  - Yes. Naive-Student to leverage Semi-supervised and Self-Supervised learning!

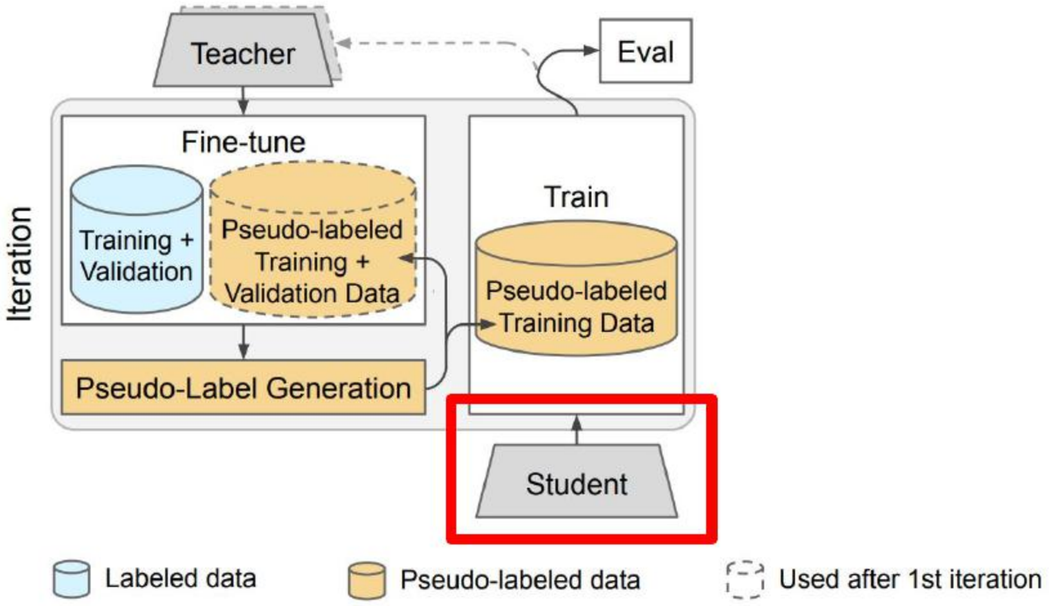
Naive-Student

- Overview:
  - Iterative semi-supervised learning with Panoptic-DeepLab.



Naive-Student: Semi-supervised learning with Panoptic-DeepLab

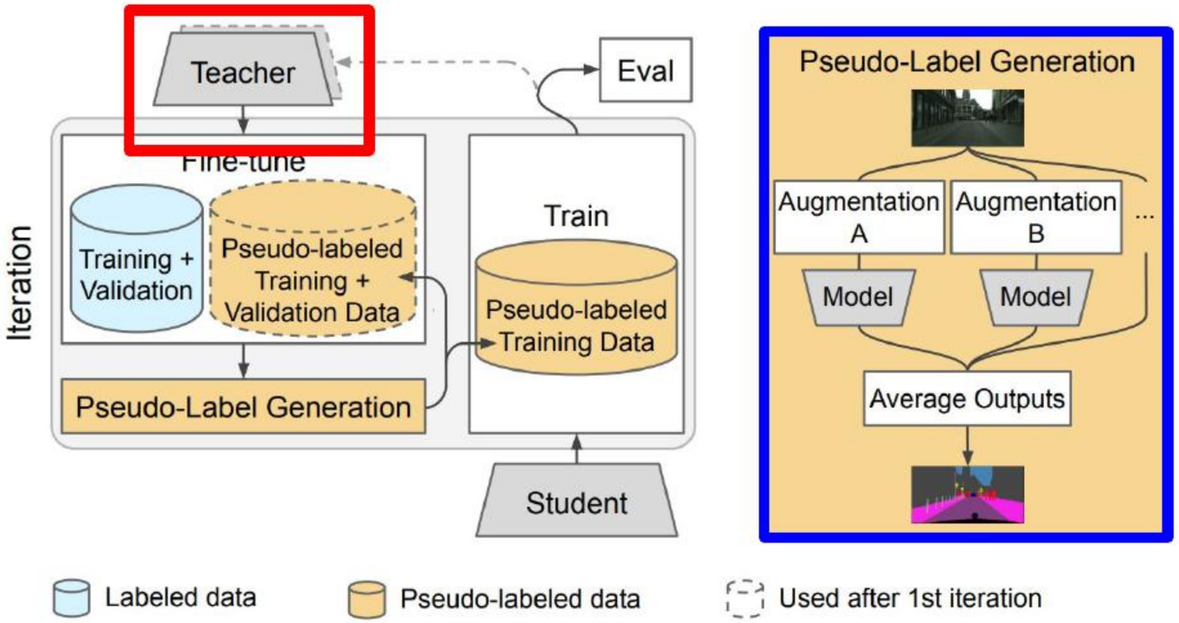
- Key components:
  a. A Student network is trained **without** injecting any noise (e.g., RandAugment).



- Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical data augmentation with no separate search. arXiv:1909.13719 (2019)

Naive-Student: Semi-supervised learning with Panoptic-DeepLab

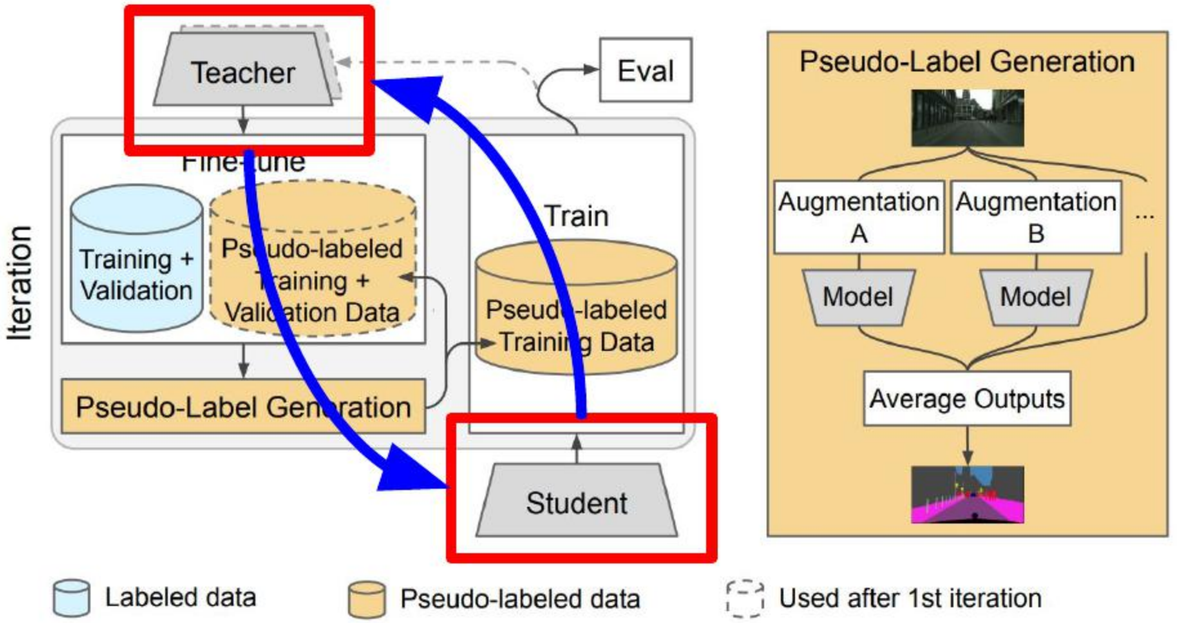
- Key components:
  a. A Student network is trained without injecting any noise (e.g., RandAugment).
  b. A Teacher network generates **pseudo-labels** with a typical test-time augmentation for each video frame or image **without** using any complicated modules.
     - No construction of sophisticated architectures.



- Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: ICML Workshop (2013)
- Mustikovela, S.K., Yang, M.Y., Rother, C.: Can ground truth label propagation from video help semantic segmentation? In: ECCV (2016)
- Radosavovic, I., Doll´ar, P., Girshick, R., Gkioxari, G., He, K.: Data distillation: Towards omni-supervised learning. In: CVPR (2018)
- Zhu, Y., Sapra, K., Reda, F.A., Shih, K.J., Newsam, S., Tao, A., Catanzaro, B.: Improving semantic segmentation via video propagation and label relaxation. In: CVPR (2019)

Naive-Student: Semi-supervised learning with Panoptic-DeepLab

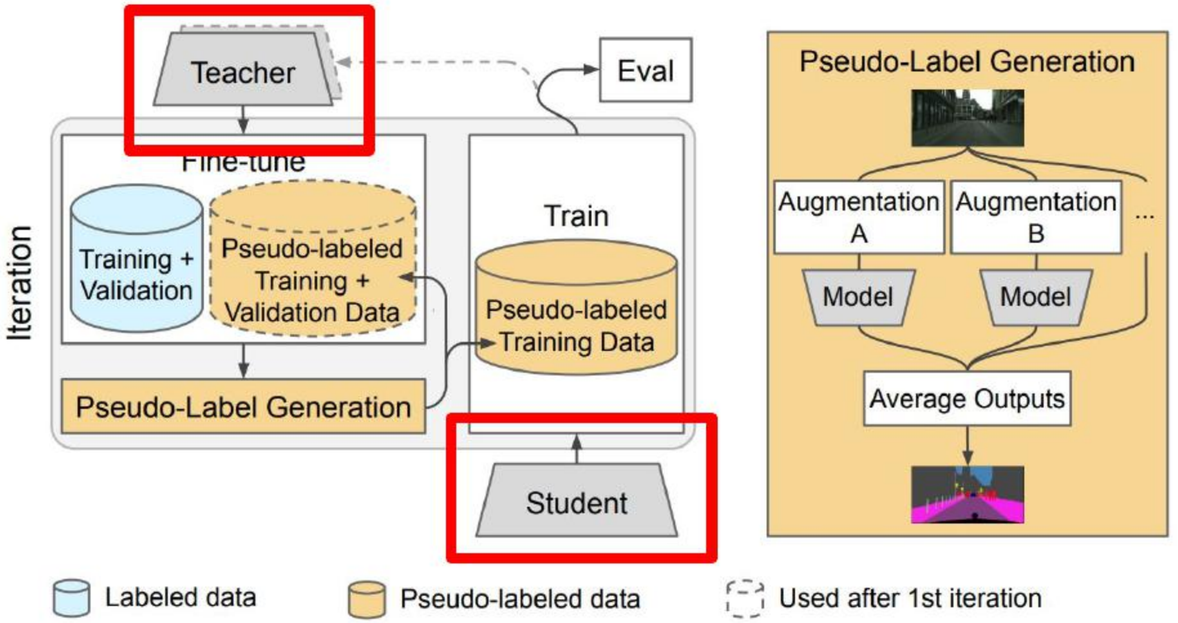
- Key components:
  a. A Student network is trained without injecting any noise (e.g., RandAugment).
  b. A Teacher network generates pseudo-labels with a typical test-time augmentation for each video frame or image without using any complicated modules.
  c. The whole process is **iterated** by employing a **stronger** student in the subsequent iterations.



- Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: ICCV (2015)
- Xie, Q., Hovy, E., Luong, M.T., Le, Q.V.: Self-training with noisy student improves imagenet classification. CVPR 2020.

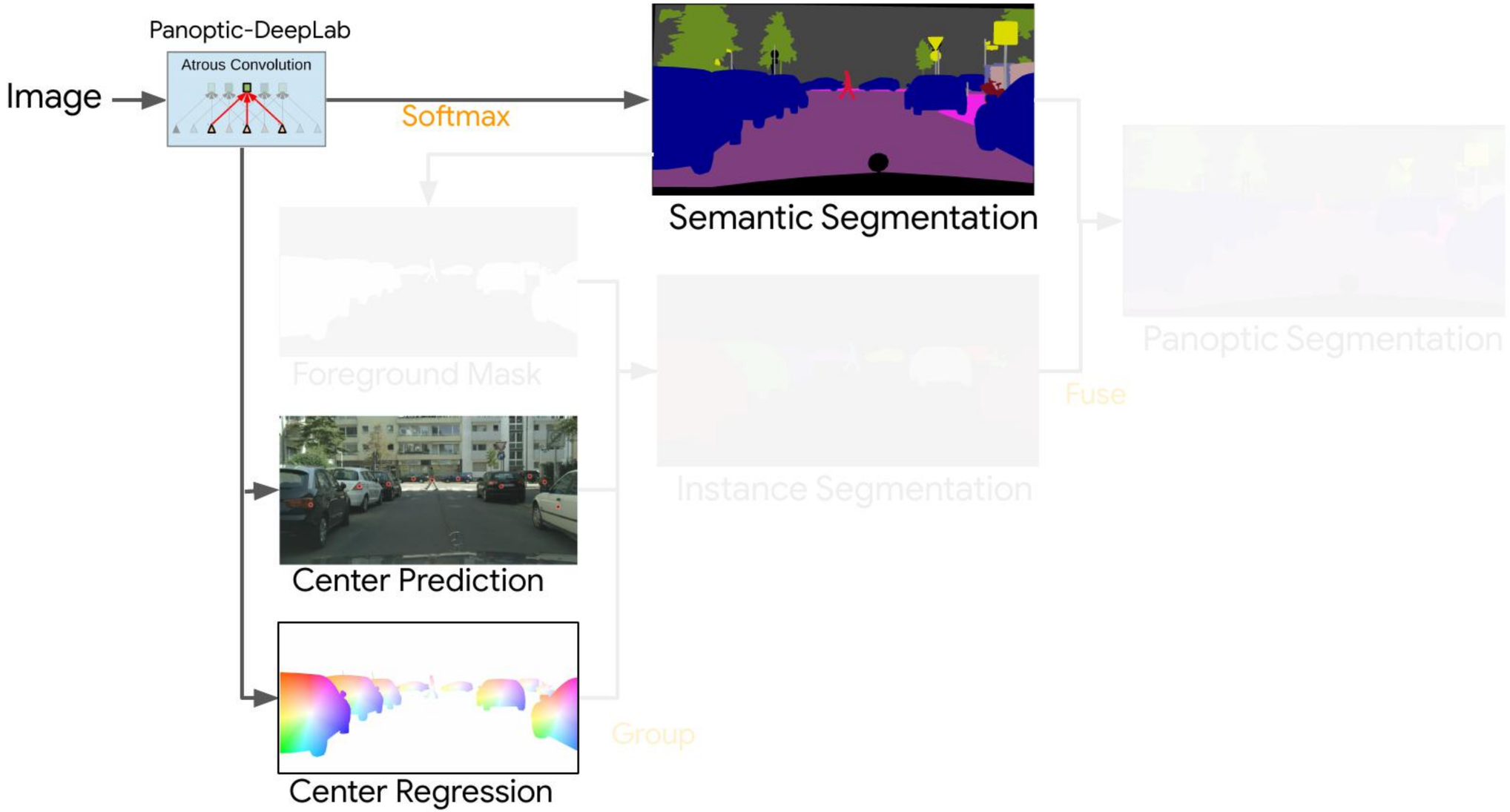
Naive-Student: Semi-supervised learning with Panoptic-DeepLab

- Key components:
  a. A Student network is trained without injecting any noise (e.g., RandAugment).
  b. A Teacher network generates pseudo-labels with a typical test-time augmentation for each video frame or image without using any complicated modules.
  c. The whole process is iterated by employing a stronger student in the subsequent iterations.
  d. **Panoptic-DeepLab** is employed for both the Teacher and Student networks.



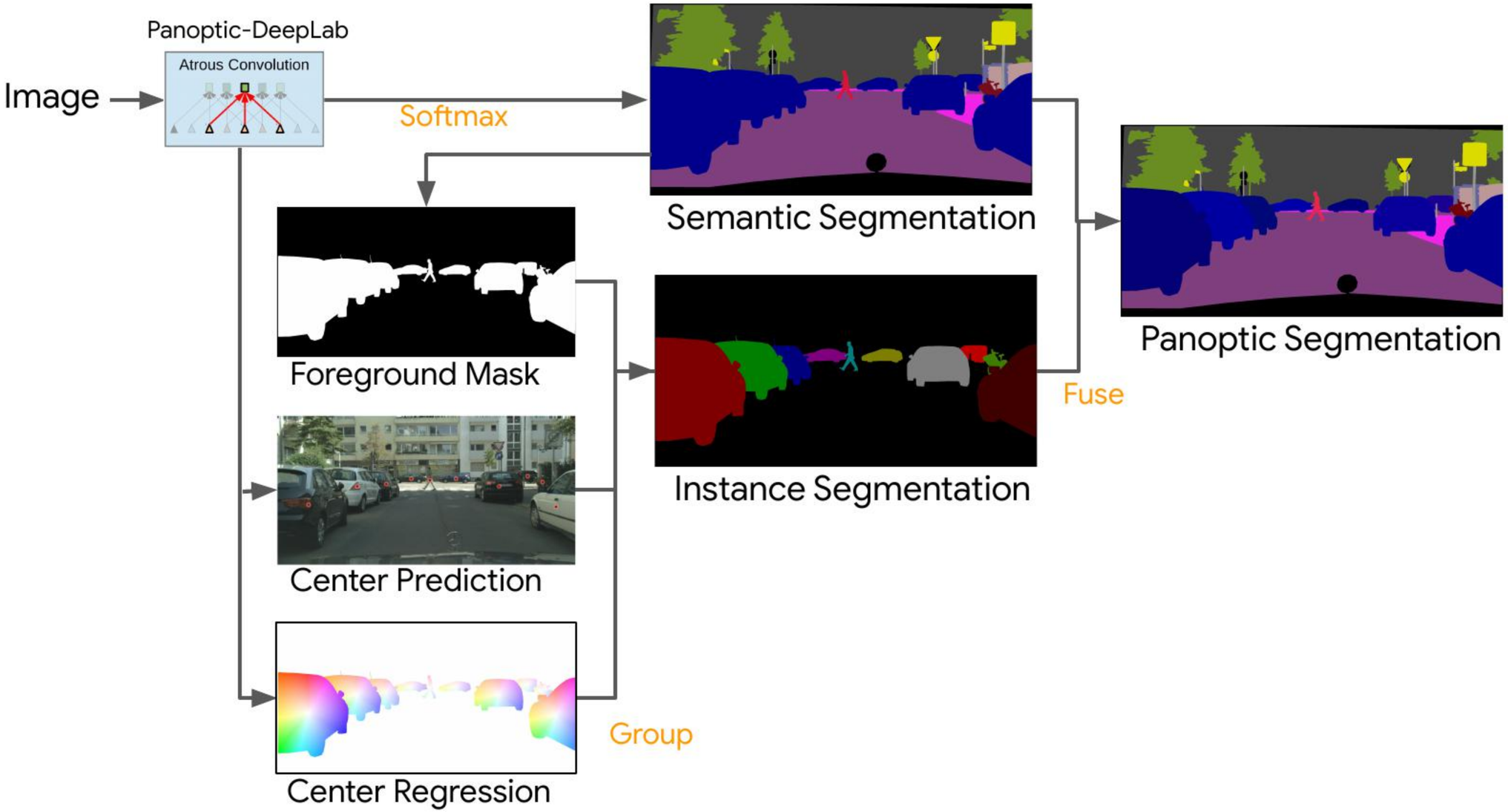
- Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.C.: Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In: CVPR (2020)

Panoptic-DeepLab: A Simple, Fast, and Strong Model



- Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.C.: Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In: CVPR (2020)

Panoptic-DeepLab: A Simple, Fast, and Strong Model



- Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.C.: Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In: CVPR (2020)

Naive-Student: Iterative Semi-supervised learning with Panoptic-DeepLab

- Algorithm:
  - At iteration 0, a Student network is trained with the small human-labeled dataset.

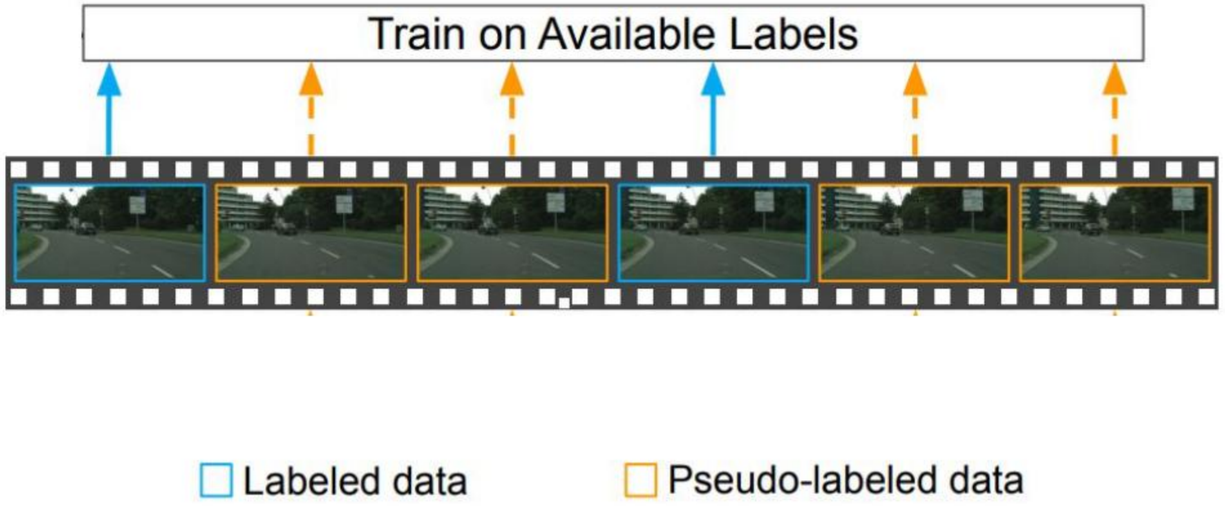
Naive-Student: Iterative Semi-supervised learning with Panoptic-DeepLab

- Algorithm:
  - At iteration 1, a Teacher network is replaced by the Student network, and generates pseudo-labels for unlabeled frames/images.

- Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: ICML Workshop (2013)
- Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: ICCV (2015)

Naive-Student: Iterative Semi-supervised learning with Panoptic-DeepLab

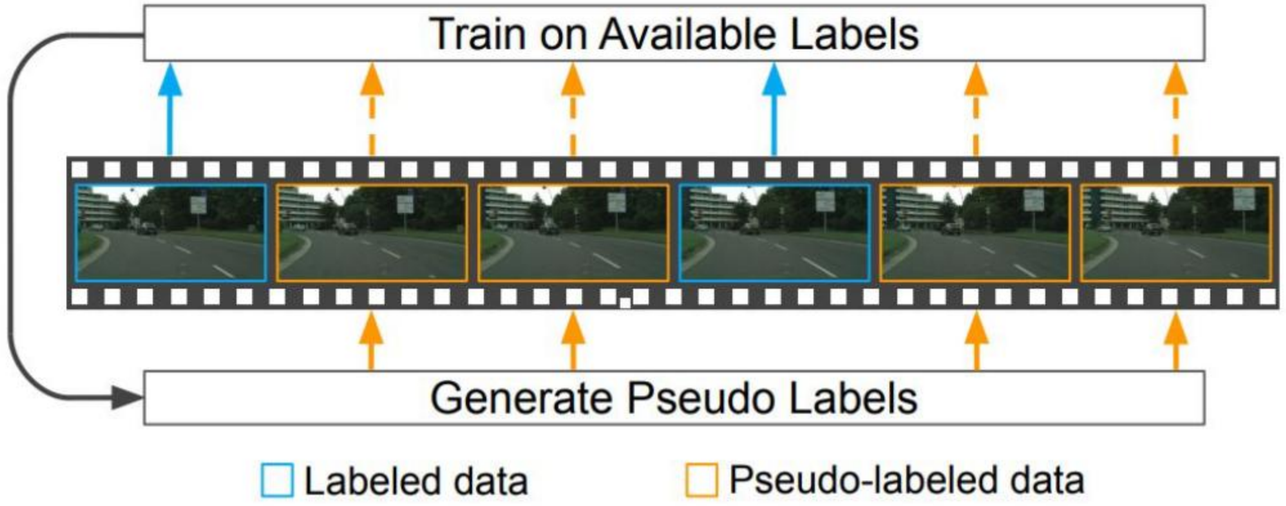
- Algorithm:
  - At iteration 1, another stronger Student is then trained on both human-annotated and pseudo-labeled data.



- Xie, Q., Hovy, E., Luong, M.T., Le, Q.V.: Self-training with noisy student improves imagenet classification. CVPR 2020.

Naive-Student: Iterative Semi-supervised learning with Panoptic-DeepLab

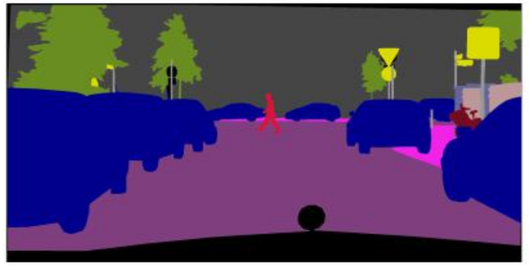
- Algorithm:
  - The process is repeated until the desired number of iterations is reached.



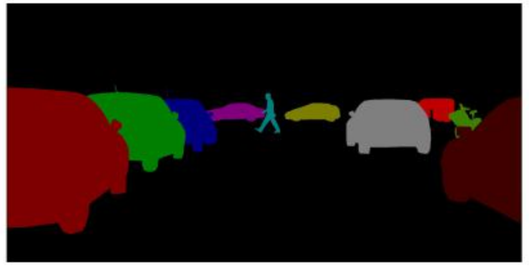
Cityscapes evaluation

- Evaluated at Cityscapes all three tasks with a single model (i.e., no fine-tuning on any specific task).
  - semantic segmentation (metric: mIOU).
    - Assign a semantic label to every pixel in an image.
  - instance segmentation (metric: AP).
    - Detect and segment each object instance.
  - panoptic segmentation (metric: PQ).
    - Joint semantic and instance segmentation.

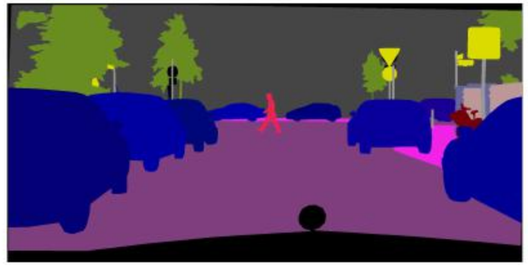
Image



Semantic Segmentation



Instance Segmentation

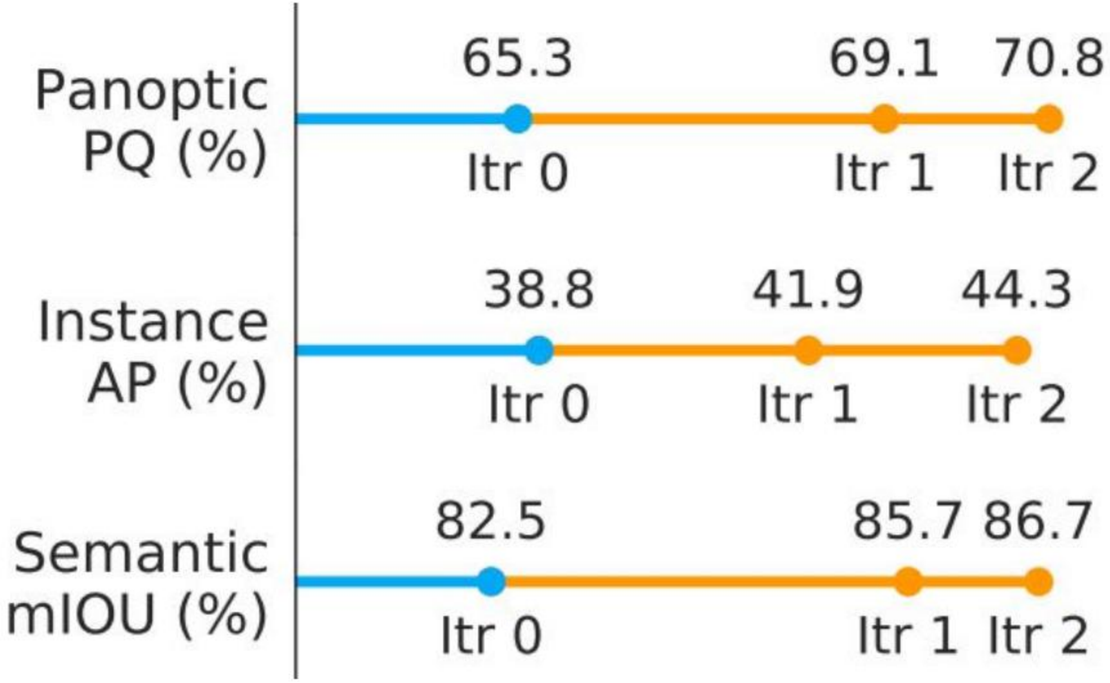


Panoptic Segmentation

- Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
- Kirillov, A., He, K., Girshick, R., Rother, C., Doll´ar, P.: Panoptic segmentation. In: CVPR (2019)

Cityscapes validation set results

- Performance improves for each iteration:



Cityscapes **test** set results

- Current state-of-the-art models:

Model	Extra Data	Training Method	PQ (%)	AP (%)	mIOU (%)
Seamless	MV	supervised	62.6	-	-
Li *et al.*	COCO	supervised	63.3	-	-
Panoptic-DeepLab (X-71)	MV	supervised	65.5	39.0	84.2
PANet	COCO	supervised	-	36.4	-
PolyTransform	COCO	supervised	-	40.1	-
Zhu *et al.*	C, V, MV	semi-supervised	-	-	83.5
OCR	C, MV	supervised	-	-	83.7

- Neuhold, G., Ollmann, T., Bul`o, S.R., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: ICCV (2017)
- Porzi, L., Bul`o, S.R., Colovic, A., Kontschieder, P.: Seamless scene segmentation. In: CVPR (2019)
- Li, Q., Qi, X., Torr, P.H.: Unifying training and inference for panoptic segmentation. arXiv:2001.04982 (2020)
- Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.C.: Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In: CVPR (2020)
- Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: CVPR (2018)
- Liang, J., Homayounfar, N., Ma, W.C., Xiong, Y., Hu, R., Urtasun, R.: Polytransform: Deep polygon transformer for instance segmentation. In: CVPR (2020)
- Zhu, Y., Sapra, K., Reda, F.A., Shih, K.J., Newsam, S., Tao, A., Catanzaro, B.: Improving semantic segmentation via video propagation and label relaxation. In: CVPR (2019)
- Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. arXiv:1909.11065 (2019)

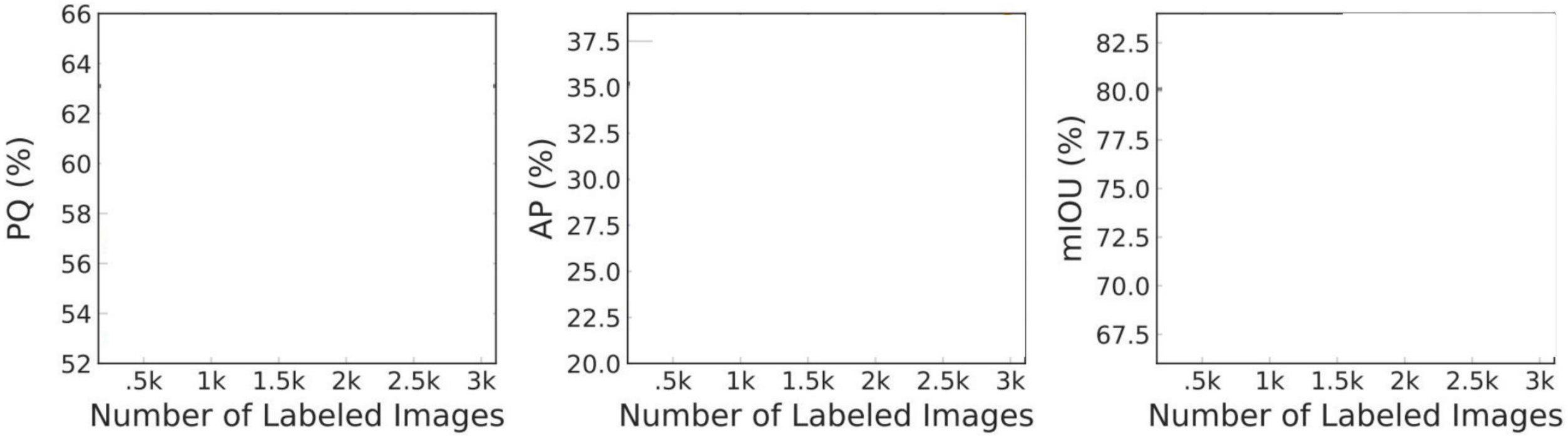
Cityscapes **test** set results

- Our method ranks 1st at all three tasks:

Model	Extra Data	Training Method	PQ (%)	AP (%)	mIOU (%)
Naive-Student (ours)	C, V, MV	iterative semi-supervised	67.8	42.6	85.2
Seamless	MV	supervised	62.6	-	-
Li *et al.*	COCO	supervised	63.3	-	-
Panoptic-DeepLab (X-71)	MV	supervised	65.5	39.0	84.2
PANet	COCO	supervised	-	36.4	-
PolyTransform	COCO	supervised	-	40.1	-
Zhu *et al.*	C, V, MV	semi-supervised	-	-	83.5
OCR	C, MV	supervised	-	-	83.7

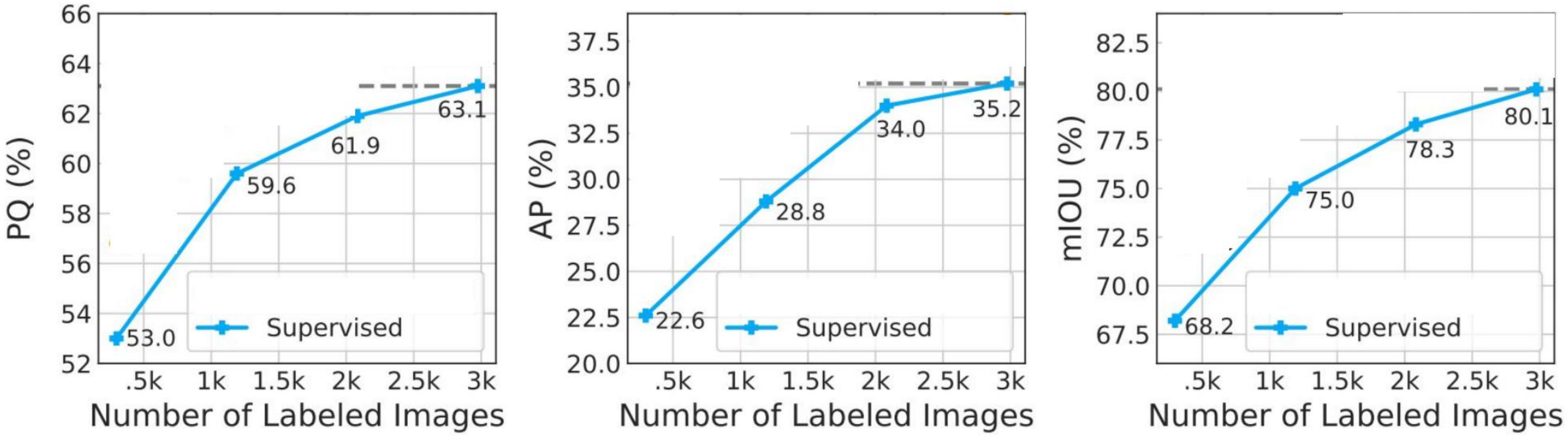
- Neuhold, G., Ollmann, T., Bul`o, S.R., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: ICCV (2017)
- Porzi, L., Bul`o, S.R., Colovic, A., Kontschieder, P.: Seamless scene segmentation. In: CVPR (2019)
- Li, Q., Qi, X., Torr, P.H.: Unifying training and inference for panoptic segmentation. arXiv:2001.04982 (2020)
- Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.C.: Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In: CVPR (2020)
- Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: CVPR (2018)
- Liang, J., Homayounfar, N., Ma, W.C., Xiong, Y., Hu, R., Urtasun, R.: Polytransform: Deep polygon transformer for instance segmentation. In: CVPR (2020)
- Zhu, Y., Sapra, K., Reda, F.A., Shih, K.J., Newsam, S., Tao, A., Catanzaro, B.: Improving semantic segmentation via video propagation and label relaxation. In: CVPR (2019)
- Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. arXiv:1909.11065 (2019)

Vary human-labeled and fix pseudo-labeled images



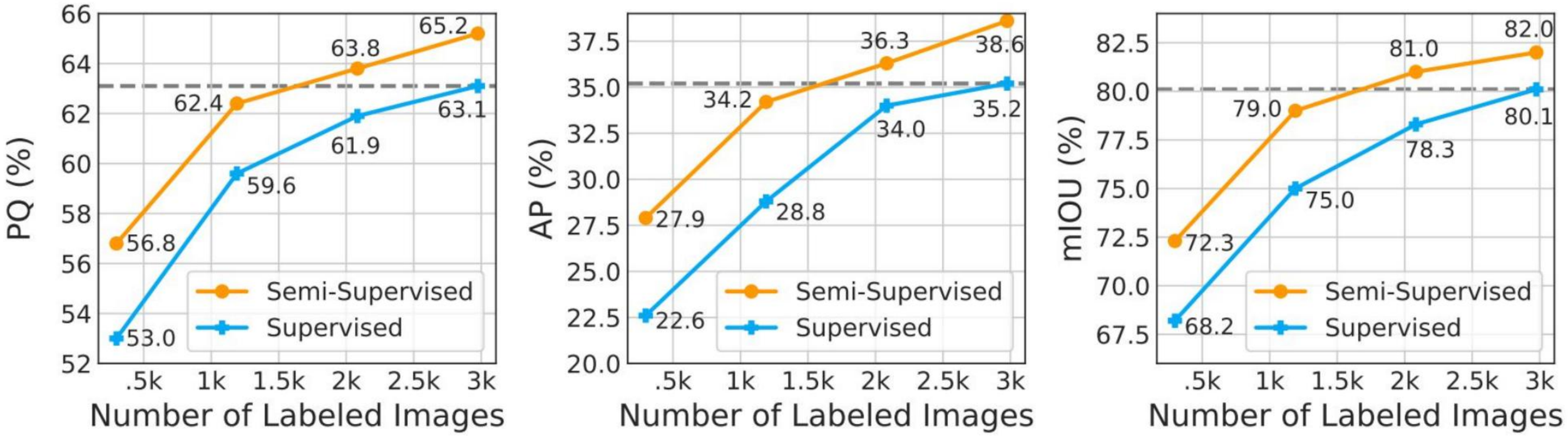
(a) PQ (b) AP (c) mIOU

Vary human-labeled and fix pseudo-labeled images



(a) PQ (b) AP (c) mIOU

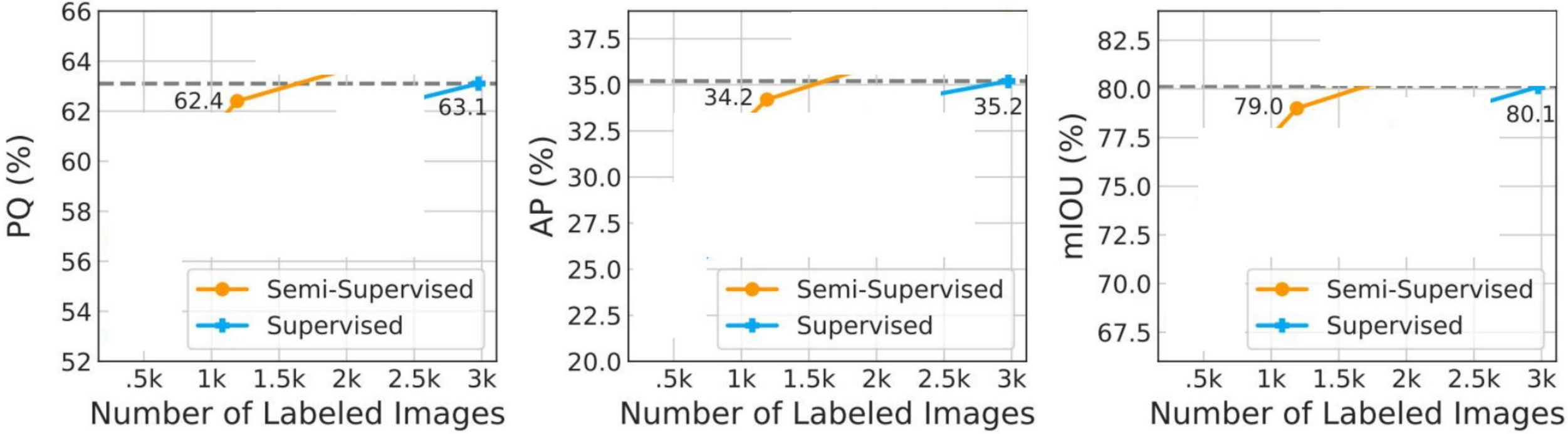
Vary human-labeled and fix pseudo-labeled images



(a) PQ (b) AP (c) mIOU

Vary human-labeled and fix pseudo-labeled images

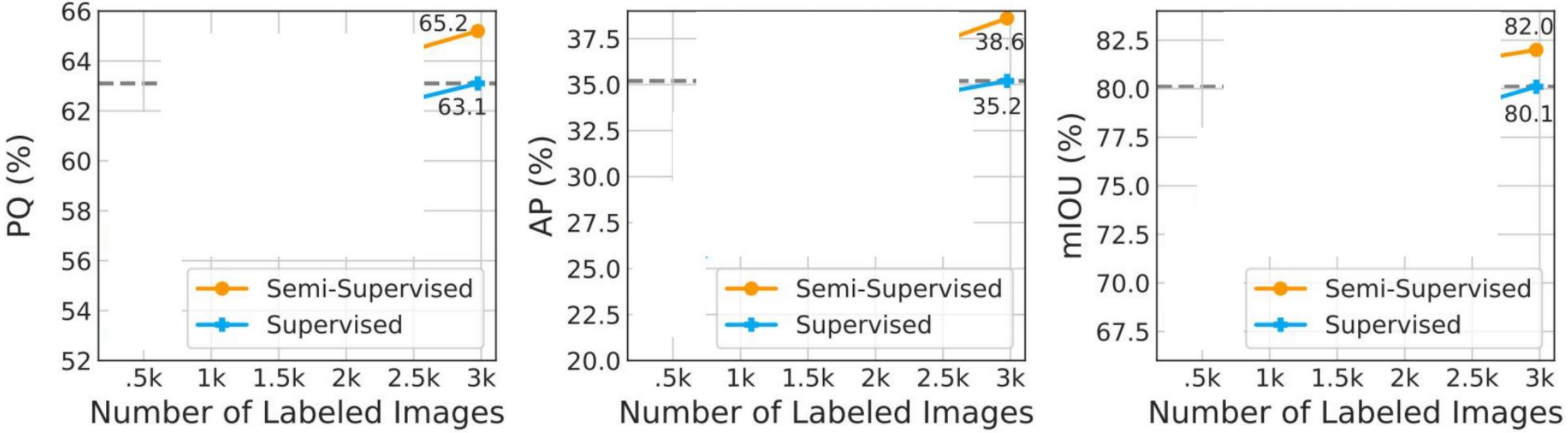
- With only **40%** of labeled data, our semi-supervised learning method could reap 98.9%, 97.2%, and 98.6% performance from its fully supervised counterparts in PQ, AP, and mIOU, respectively.



(a) PQ (b) AP (c) mIOU

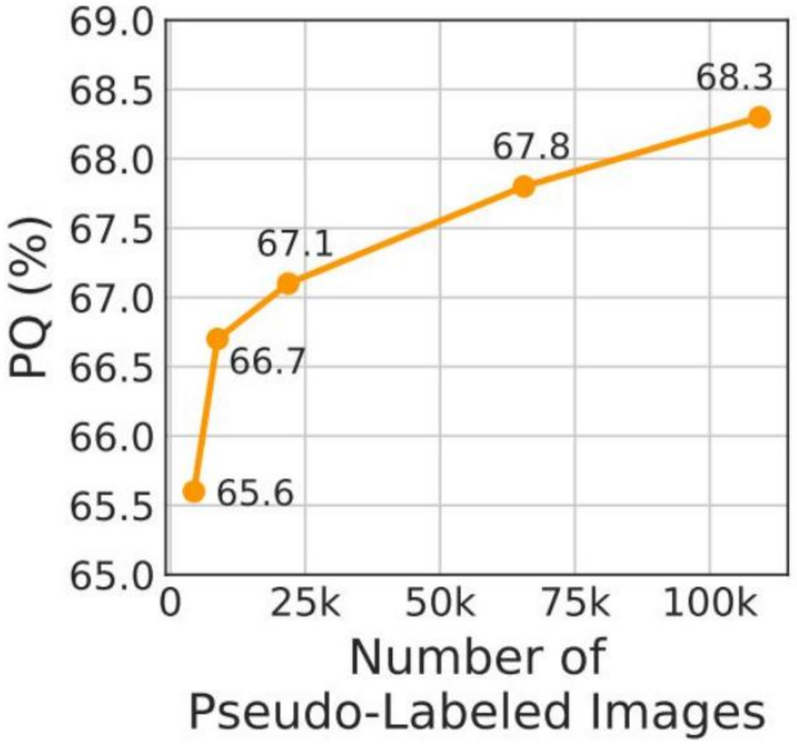
Vary human-labeled and fix pseudo-labeled images

- When using **100%** of the human-labeled images, our semi-supervised learning method attains 65.2% PQ, 38.6% AP, and 82% mIOU, comparable to fully supervised counterpart with a Mapillary Vistas pretrained checkpoint.

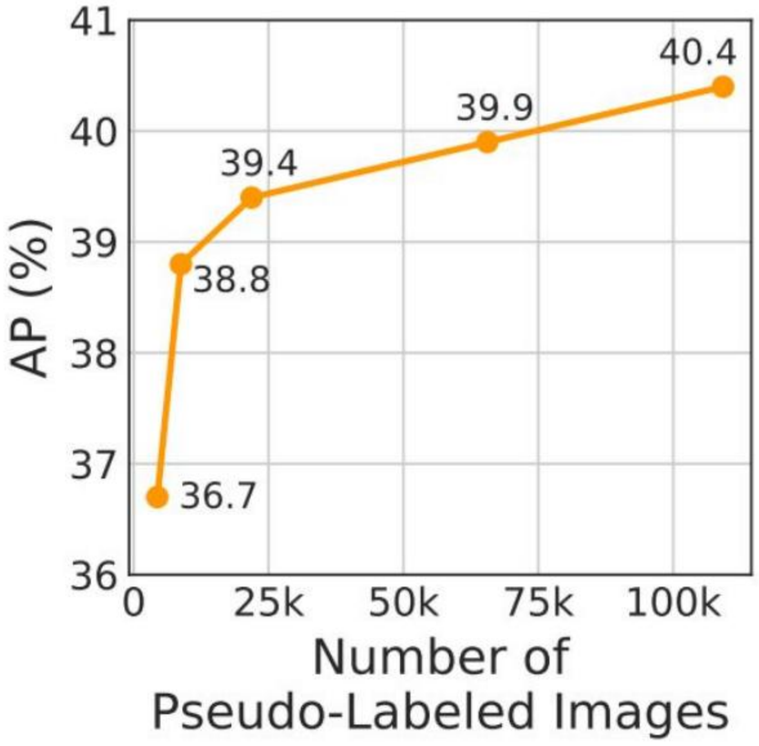


(a) PQ (b) AP (c) mIOU

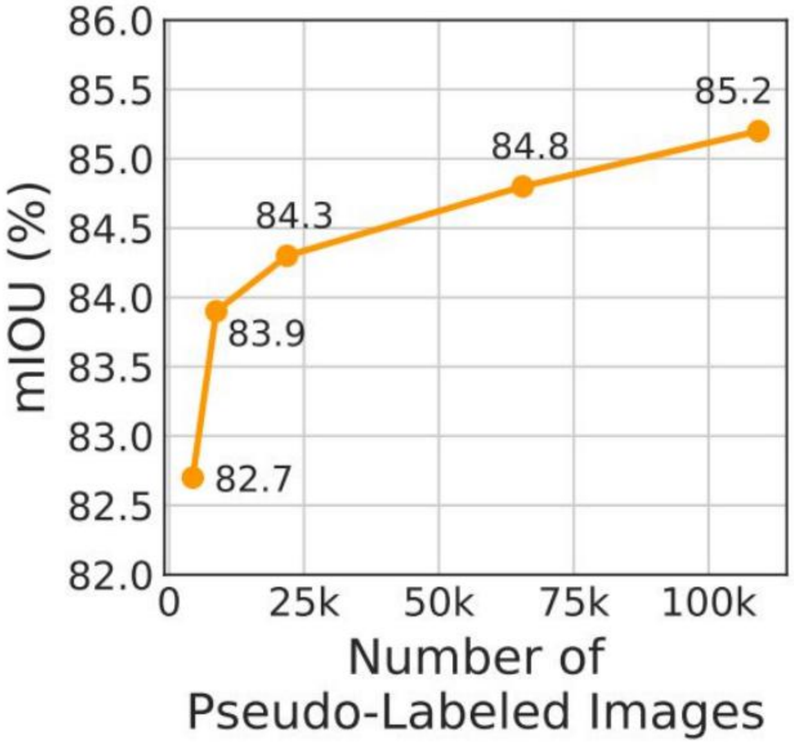
Fix human-labeled and Vary pseudo-labeled images



(a) PQ



(b) AP



(c) mIOU

Conclusion

- We propose Naive-Student, a simple iterative semi-supervised learning method for scene segmentation, **simultaneously** tackling
  - semantic segmentation
  - instance segmentation
  - panoptic segmentation

Conclusion

- We propose Naive-Student, a simple iterative semi-supervised learning method for scene segmentation, **simultaneously** tackling
  - semantic segmentation
  - instance segmentation
  - panoptic segmentation
- Our method effectively exploits both **unlabeled** video frames and extra images to improve the performance **without** the creation of additional architectures and learned modules.