A baseline for semi-supervised learning of efficient semantic segmentation models

Ivan Grubišić, Marin Oršić, Siniša Šegvić

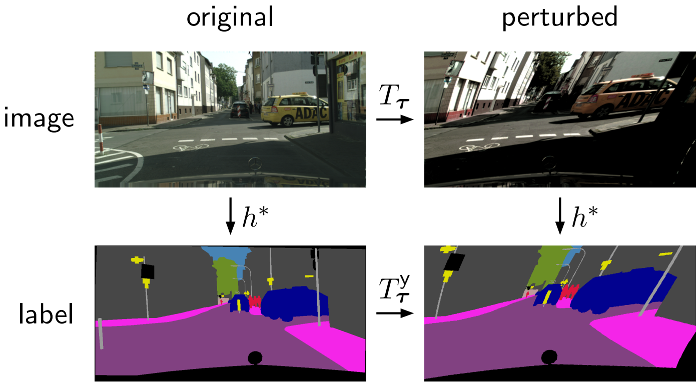
Faculty of Electrical Engineering and Computing
Department of Electronics, Microelectronics, Computer and Intelligent Systems

Overview

- Semi-supervised learning (SSL) is interesting in the **dense prediction** context.
- Evaluation on an **efficient** architecture,
- Enforcement of **one-way consistency** under **photometric and geometric input perturbations**.
- We investigate some consistency training choices.

SSL with input perturbation consistency

- Enforcement of **prediction consistency** under **different input perturbations** or **different model instances**.
- Some perturbations are such that the correct output is not invariant to them.



One-way consistency

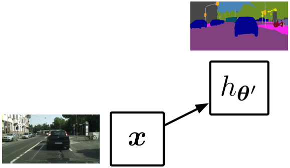
- The **student** $h_{\boldsymbol{\theta}}$, is trained to be consistent with the **teacher** $h_{\boldsymbol{\theta}'}$.
- The simplest algorithm,
  - $\boldsymbol{\theta}'$ is an independent copy of $\boldsymbol{\theta}$ and
  - the student's input is perturbed.
- Alternative teacher: Mean Teacher pseudo-ensembling [3].
- The memory footprint of supervised training.

One-way consistency

- Let $\boldsymbol{x}$ be the unlabeled input, $h_{\boldsymbol{\theta}}$ the student, $h_{\boldsymbol{\theta}'}$ the teacher, $T_{\boldsymbol{\tau}}$ and $T_{\boldsymbol{\tau}}^{\mathsf{y}}$ the corresponding input and output perturbations, and D a measure of distance between two distributions.
- Only the blue part of the graph is used for computing the gradient.

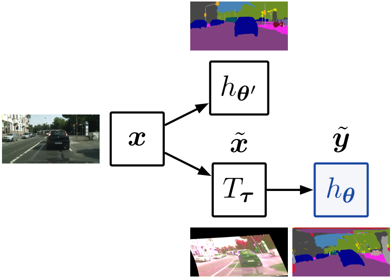
One-way consistency

- Let $\boldsymbol{x}$ be the unlabeled input, $h_{\boldsymbol{\theta}}$ the student, $h_{\boldsymbol{\theta}'}$ the teacher, $T_{\boldsymbol{\tau}}$ and $T_{\boldsymbol{\tau}}^{\mathsf{y}}$ the corresponding input and output perturbations, and D a measure of distance between two distributions.
- Only the blue part of the graph is used for computing the gradient.



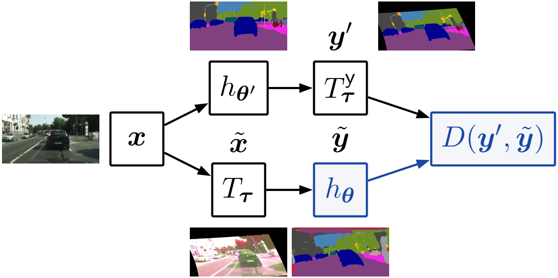
One-way consistency

- Let $\boldsymbol{x}$ be the unlabeled input, $h_{\boldsymbol{\theta}}$ the student, $h_{\boldsymbol{\theta}'}$ the teacher, $T_{\boldsymbol{\tau}}$ and $T_{\boldsymbol{\tau}}^{\mathsf{y}}$ the corresponding input and output perturbations, and D a measure of distance between two distributions.
- Only the blue part of the graph is used for computing the gradient.



One-way consistency

- Let $\boldsymbol{x}$ be the unlabeled input, $h_{\boldsymbol{\theta}}$ the student, $h_{\boldsymbol{\theta}'}$ the teacher, $T_{\boldsymbol{\tau}}$ and $T_{\boldsymbol{\tau}}^{\mathsf{y}}$ the corresponding input and output perturbations, and D a measure of distance between two distributions.
- Only the blue part of the graph is used for computing the gradient.



Our method

- We achieve best results with Mean Teacher (MT) and our perturbation model (PhTPS) – a composition of a photometric and a geometric transformation.

Our method

- We express our consistency loss as mean per-pixel KL-divergence over valid prediction pixels.
- Since the gradient is not propagated through the teacher and

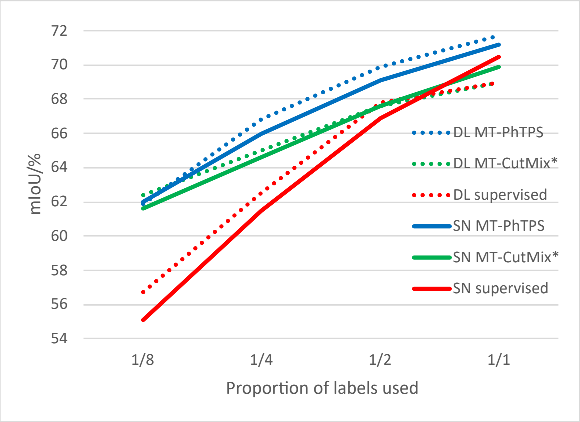
$$D(\underline{y}, \tilde{\underline{y}}) = \mathop{\mathbf{E}}_{\underline{y}} \ln \frac{\mathrm{P}(\underline{y} = y)}{\mathrm{P}(\tilde{\underline{y}} = y)} = \mathrm{H}_{\tilde{\underline{y}}}(\underline{y}) - \mathrm{H}(\underline{y}),$$

  the entropy increasing term $-\mathrm{H}(\underline{y})$ has no effect on parameter updates; only the cross-entropy term has an effect.

Experiments

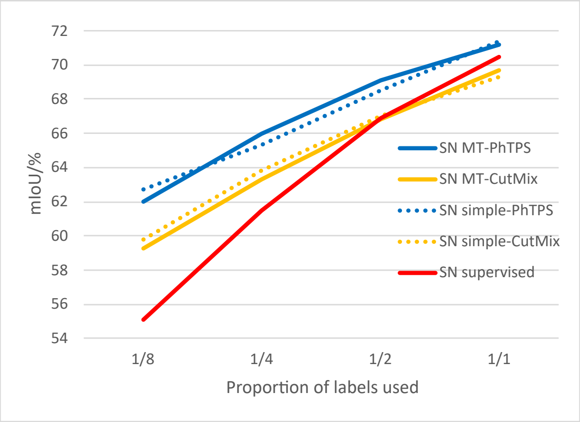
- For the efficient model, we use SwiftNet-RN18 (**SN**), which is $\sim 5\times$ faster to train in our configurations and $\sim 12\times$ faster to evaluate than DeepLabv2 (**DL**).
- The compared SSL algorithm configurations:
  - The teacher can equal the student (**simple** consistency) or be a "mean teacher" (**MT**).
  - The perturbations can be ours (**PhTPS**) or **CutMix**. We also test CutMix with L2 loss and confidence thresholding [1] (**CutMix***).

Half-resolution Cityscapes label subsets



- SwiftNet-RN18 (solid) is slightly worse than DeepLabv2 (dotted).
- The models behave similarly: PhTPS $\succ$ CutMix* $\succ$ supervised.

Half-resolution Cityscapes label subsets



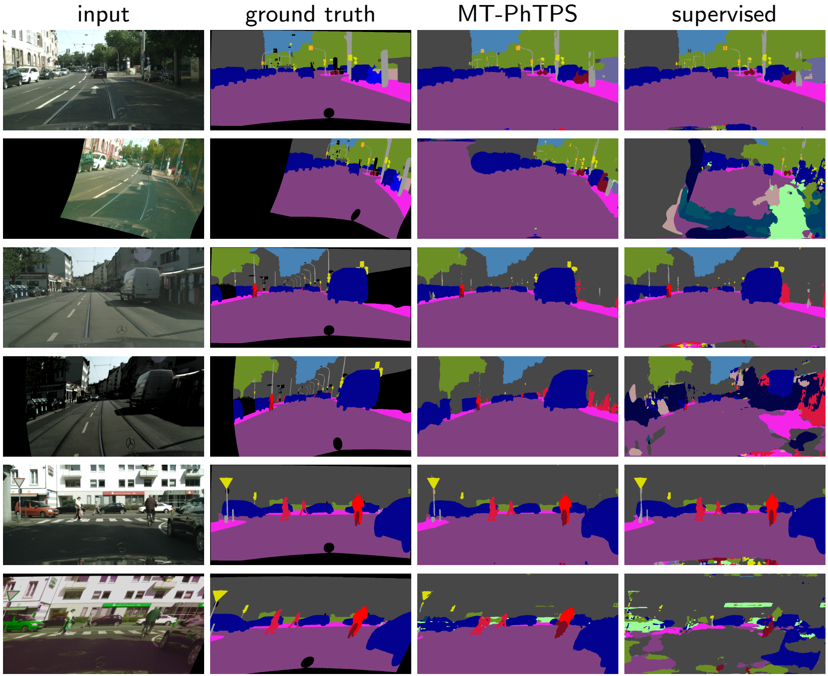
- Comparison of SSL configurations under SwiftNet-RN18.
- Mostly PhTPS $\succ$ CutMix $\succ$ supervised (maybe MT $\succ$ simple).

Comparison of consistency variants

- We compare consistency variants on CIFAR-10 classification and half-resolution Cityscapes semantic segmentation.
  - "1w" denotes one-way consistency. "ps", "pt", and "p2" denote perturbation of the teacher's, the student's, and both inputs.
  - "2w-p1" denotes two-way consistency with 1 input perturbed.

Dataset	SSL algorithm	sup.	1w-ps	1w-pt	2w-p1	1w-p2
CIFAR-10, 2/25	simple-PhTPS	$80.8_{0.4}$	$\mathbf{90.8}_{0.3}$	$50.1_{20.1}$	$72.9_{1.0}$	$73.3_{7.0}$
CIFAR-10, 2/25	MT-PhTPS	$80.8_{0.4}$	$\mathbf{90.8}_{0.4}$	$80.5_{0.5}$	-	$73.4_{1.4}$
Cityscapes, 1/4	simple-PhTPS	$61.5_{0.5}$	$\mathbf{65.3}_{1.9}$	$1.6_{1.0}$	$16.7_{3.0}$	$61.6_{0.5}$
Cityscapes, 1/4	MT-PhTPS	$61.5_{0.5}$	$\mathbf{66.0}_{1.0}$	$61.5_{1.4}$	-	$62.0_{1.1}$

- Only 1w-ps significantly outperforms the supervised baseline.


input
ground truth
MT-PhTPS
supervised

Conclusion

- It might be good to consider efficient models for comparison of semi-supervised semantic segmentation algorithms ($\sim 5\times$ faster training, $\sim 12\times$ faster inference).
- Our perturbation model (PhTPS) outperformed CutMix.
- Mean Teacher slightly outperformed simple consistency with our perturbations.
- One-way consistency with a perturbed student outperformed all alternative consistency variants.

References

[1] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson.
Semi-supervised semantic segmentation needs strong, varied perturbations.
In *BMVC*, 2020.

[2] T. Miyato, S. Maeda, M. Koyama, and S. Ishii.
Virtual adversarial training: A regularization method for supervised and semi-supervised learning.
IEEE Trans. Pattern Anal. Mach. Intell., 41(8):1979–1993, 2019.

[3] A. Tarvainen and H. Valpola.
Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.
In *Advances in neural information processing systems*, pages 1195–1204, 2017.

[4] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le.
Unsupervised data augmentation for consistency training.
In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6256–6268. Curran Associates, Inc., 2020.