17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

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

Ivan Grubišić*, Marin Oršić*, Siniša Šegvić
University of Zagreb, Faculty of Electrical Engineering and Computing
ivan.grubisic@fer.hr, marin.orsic@gmail.com, sinisa.segvic@fer.hr

Abstract

Semi-supervised learning is especially interesting in the dense prediction context due to high cost of pixel-level ground truth. Unfortunately, most such approaches are evaluated on outdated architectures which hamper research due to very slow training and high requirements on GPU RAM. We address this concern by presenting a simple and effective baseline which works very well both on standard and efficient architectures. Our baseline is based on one-way consistency and non-linear geometric and photometric perturbations. We show advantage of perturbing only the student branch and present a plausible explanation of such behaviour. Experiments on Cityscapes and CIFAR-10 demonstrate competitive performance with respect to prior work.

1 Introduction

Semi-supervised learning [1, 2] presents a great opportunity to speed-up the development cycle and enable rapid adaptation to new environments. It is especially relevant in the dense prediction context [3, 4, 5, 6] since pixel-level labels are very expensive.

Many semi-supervised approaches are based on enforcing consistent predictions in differently perturbed inputs [1, 2, 7, 8, 9]. Perturbations can be random [1], adversarial [8], geometric [7], or even non-differentiable [6]. The learning signal can be improved by averaging predictions [1] or model parameters [2, 6]. Some approaches use one-way consistency, which allows the gradients to pass only through the student branch, while the teacher branch is frozen [2, 8, 9, 6]. Some of these works perturb only the student branch [8, 9, 6]; however, none of them discuss advantages of that setup.

In practice, semi-supervised algorithms train on all available supervised data, while also exploiting a much larger quantity of unsupervised data. The computational strain is especially large in the dense-prediction case since many practical applications require large input resolutions [10, 11, 12]. Many semi-supervised algorithms further increase the memory footprint of training (training footprint) due to extra logits [13], a GAN generator [14, 3] or discriminator [4, 15], or multiple model instances [16, 7, 17]. Such designs are less appropriate for practical dense prediction since their training footprint constrains the backbone complexity [18] and requires expensive hardware. Unfortunately, current research mostly involves inefficient models [19] which require a lot of GPU RAM and training time.

We propose a simple and effective method for semi-supervised semantic segmentation. Our method is based on one-way consistency [8, 9, 6] and non-linear perturbations which outperform a recent baseline [6]. One-way consistency is advantageous since it retains the training footprint of the underlying supervised algorithm, unlike [4, 20, 5, 21]. Additionally, it outperforms two-way consistency in terms of generalization performance, which we demonstrate by presenting intuitive arguments and empirical evidence.

Experiments with the standard convolutional architecture [19] reveal competitive accuracy. We observe a similar advantage in experiments with a recent efficient architecture [22], which performs close while requiring an order of magnitude less computation. This is the first account of evaluation of semi-supervised algorithms for dense prediction with a model capable of real-time inference. This contributes to the goals of Green AI [23] by enabling relevant research on inexpensive hardware while reducing environmental damage.

2 Related Work

A semantic segmentation model can leverage unlabeled images through GAN training as a dense discriminator [3]. KE-GAN [21] additionally enforces semantic consistency of neighbouring predictions by leveraging label-similarity recovered from a large text corpus (MIT ConceptNet). A semantic segmentation model can also be trained as a GAN generator [4] in order to encourage more realistic predictions. s4GAN [5] additionally post-processes dense predictions by removing classes not identified by an image-wide classifier trained with Mean Teacher [2]. Universal semantic segmentation [20] pulls features from unlabeled images towards centroids obtained by training on multiple datasets.

Dense semi-supervision can also be based on pseudo-labels [24, 25, 26, 27]. This can be improved by iterative noisy-student training [26]. Mixing pseudo-labeled with labeled images reduces performance in this setup.

A recent approach [7] proposes Π-style [1] two-way consistency over geometric warps. Another recent approach [28] enforces consistency between outputs of

[*] Equal contribution.

redundant decoders with noisy intermediate representations. Mean Teacher consistency with CutMix perturbations [6] obtains state-of-the-art performance on half-resolution Cityscapes.

Our method is related to [7] who also perturb images with geometric warps. However, we show that perturbing only the student branch generalizes much better than two-way consistency and has a smaller training footprint. Different than most presented approaches and similar to [25, 26, 6], our method does not increase the training footprint [18]. In comparison with [25, 26], our teacher is updated in each training step, which eliminates the need for multiple training episodes. In comparison with [6], we use a perturbation model which results in better accuracy. None of the previous approaches addresses semi-supervised training of efficient dense prediction models [29, 30, 22, 31].

3 Dense One-Way Consistency

We formulate dense consistency as a mean pixel-wise divergence between corresponding predictions in the clean image and its perturbed version. We perturb images according to a composition of parametric transformations of color and shape.

3.1 Notation and preliminaries

We typeset arrays and vectors in bold, sets in blackboard bold, and random variables underlined. We use Python-like indexing notation.

We denote the labeled and the unlabeled dataset as $\mathbb{D}_\mathrm{l}$ and $\mathbb{D}_\mathrm{u}$, respectively. We consider input images $\boldsymbol{x} \in \mathbb{X} = [0,1]^{H\times W\times 3}$ and dense labels $\boldsymbol{y} \in \mathbb{Y} = \{1..C\}^{H\times W}$. A model instance maps an image to per-pixel class probabilities: $h_{\boldsymbol{\theta}}(\boldsymbol{x})_{[i,j,c]} = \mathrm{P}(\underline{\boldsymbol{y}}_{[i,j]} = c|\boldsymbol{x},\boldsymbol{\theta})$. For convenience, we identify output vectors with distributions: $h_{\boldsymbol{\theta}}(\boldsymbol{x})_{[i,j]} \equiv \mathrm{P}_{\underline{\boldsymbol{y}}_{[i,j]}|\boldsymbol{x},\boldsymbol{\theta}}$.

We consider teacher parameters $\boldsymbol{\theta}'$ as a frozen copy of either student parameters $\boldsymbol{\theta}$ (simple consistency) or their moving average (Mean Teacher). In two-way consistency $\boldsymbol{\theta}' = \boldsymbol{\theta}$.

Our perturbation $T_{\boldsymbol{\tau}} = T^\mathrm{G}_{\boldsymbol{\gamma}} \circ T^\mathrm{P}_{\boldsymbol{\varphi}}$ is a composition of a geometric warp $T^\mathrm{G}_{\boldsymbol{\gamma}}$ and a global photometric transformation $T^\mathrm{P}_{\boldsymbol{\varphi}}$ with parameters $\boldsymbol{\tau} = (\boldsymbol{\gamma}, \boldsymbol{\varphi})$. $T_{\boldsymbol{\gamma}}$ displaces pixels with a dense deformation field and internally uses zero-padding and bilinear interpolation.

3.2 One-way consistency with a clean teacher

We express a general semi-supervised training criterion as a combinaton of a supervised term L_s and an unsupervised consistency term L_c:

$$E(\boldsymbol{\theta}; \mathbb{D}_\mathrm{l}, \mathbb{D}_\mathrm{u}) = \mathop{\mathbf{E}}_{(\boldsymbol{x},\boldsymbol{y})\in\mathbb{D}_\mathrm{l}} L_\mathrm{s}(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}) + \alpha \mathop{\mathbf{E}}_{\boldsymbol{x}\in\mathbb{D}_\mathrm{u}} L_\mathrm{c}(\boldsymbol{\theta}; \boldsymbol{x}) \ . \quad (1)$$

In our case, L_c encourages predictions to be invariant under photometric perturbations $T^\mathrm{P}_{\boldsymbol{\varphi}}$ and equivariant over geometric perturbations $T^\mathrm{G}_{\boldsymbol{\gamma}}$. We can define it as the expected average of a divergence D between corresponding predictions of the teacher with the unperturbed input (clean teacher) and the student with the perturbed input (perturbed student). We compute $L_\mathrm{c}(\boldsymbol{\theta}; \boldsymbol{x})$ by sampling perturbation parameters $\boldsymbol{\tau} = (\boldsymbol{\gamma}, \boldsymbol{\varphi})$ and averaging the following per-pixel loss:

$$L^{i,j}_\mathrm{c}(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{\tau}) = D(T^\mathrm{G}_{\boldsymbol{\gamma}}(h_{\boldsymbol{\theta}'}(\boldsymbol{x}))_{[i,j]}, h_{\boldsymbol{\theta}}(T_{\boldsymbol{\tau}}(\boldsymbol{x}))_{[i,j]}) \quad (2)$$

The warped teacher prediction $T^\mathrm{G}_{\boldsymbol{\gamma}}(h_{\boldsymbol{\theta}'}(\boldsymbol{x}))_{[i,j]}$ is always a valid distribution since the aggregation ignores pixels from padding. More precisely, we aggregate the loss only in pixels (i,j) where $T^\mathrm{G}_{\boldsymbol{\gamma}}(\mathbf{1}_{H\times W})_{[i,j]} = 1$. Due to $\boldsymbol{\theta}'$ being a frozen copy, effectively, the gradient propagates only through the student branch, towards the perturbed image.

We use KL divergence as a principled choice for D. Since the gradient is not propagated through $\boldsymbol{\theta}'$ and $D(\underline{y}, \underline{\tilde{y}}) = \mathrm{H}_{\underline{\tilde{y}}}(\underline{y}) - \mathrm{H}(\underline{y})$, the entropy increasing term $-\mathrm{H}(\underline{y})$ has no effect on parameter updates; only the cross-entropy term has an effect.

Our experiments show that clean teachers generalize better than perturbed teachers. We know that the standard supervised loss promotes invariance to weak perturbations due to jittering. Hence, the consistency loss requires stronger perturbations to create more opportunity for learning from differences in predictions, as also noted in [6]. However, strong perturbations may push inputs beyond the natural manifold and spoil predictions. We observe that perturbing both branches sometimes results in learning to map perturbed inputs to similar arbitrary predictions (e.g. always the same class). Hence, consistency training has best chances to succeed when the teacher is applied to clean inputs.

3.3 Photometric and geometric perturbations

Our photometric perturbations are a composition of five pixel-level transformations with image-wide parameters $\boldsymbol{\varphi} = (b, s, h, c, \boldsymbol{\pi})$. The compound perturbation $T^\mathrm{P}_{\boldsymbol{\varphi}}$ can be described as follows: (1) brightness is shifted by adding b to all channels, (2) saturation is multiplied with s, (3) hue is shifted by addition with h, (4) contrast is modulated by multiplying all channels with c, and (5) RGB channels are permuted according to $\boldsymbol{\pi}$. $\boldsymbol{\varphi}$ is randomly picked: $b \sim \mathcal{U}(-0.25, 0.25)$, $s \sim \mathcal{U}(0.25, 2)$, $h \sim \mathcal{U}(-36^\circ, 36^\circ)$, $c \sim \mathcal{U}(0.25, 2)$.

We formulate a class of parametric geometric transformations by leveraging thin plate splines (TPS) [32, 33] which transform a 2D point $\boldsymbol{q}$ as follows:

$$f(\boldsymbol{q}) = \boldsymbol{A} \cdot \begin{bmatrix} 1 \\ \boldsymbol{q} \end{bmatrix} + \boldsymbol{W} \cdot [\phi(\|\boldsymbol{q} - \boldsymbol{c}_i\|)]^\mathsf{T}_{i=1..n} \ . \quad (3)$$

In the equation, $\boldsymbol{c}_i$ are control points, $\boldsymbol{A}$ is a 2×3 affine transformation matrix, $\boldsymbol{W}$ is a 2×n control point coefficient matrix, and $\phi(r) = r^2 \ln(r)$. $\boldsymbol{A}$ and $\boldsymbol{W}$ are obtained by solving the linear system $\bigwedge_i f(\boldsymbol{c}_i) = \boldsymbol{d}_i$, where $\boldsymbol{d}_i$ are displacements of control points. Our warp $T^{\mathrm{G}}_{\boldsymbol{\gamma}}(\boldsymbol{x})$ resamples $\boldsymbol{x}$ according to f defined by displacements of four input quadrant centers: $\boldsymbol{\gamma} = (\boldsymbol{d}_1, .., \boldsymbol{d}_4)$, We sample each displacement from $\mathcal{N}(\mathbf{0}_2, r\mathbf{I}_2)$, where r is the maximum L^∞ norm of the displacement. We choose $r = 0.05 \cdot H$, where H is the image height.

4 Experiments

Our experiments evaluate generalization potential of the proposed method. We denote simple one-way consistency as "simple", Mean Teacher as "MT", and our perturbations as "PhTPS". We present means and standard deviations on 5 subsets except for experiments with DeepLab v2. Firstly, we compare semi-supervised algorithms and their components on half-resolution Cityscapes [10]. Secondly, we compare various consistency variants on Cityscapes and CIFAR-10. Source code for reproducing experiments is available at `https://github.com/Ivan1248/semisup-seg-efficient`.

4.1 Experimental setup

We perform semantic segmentation on Cityscapes [10] and image classification on CIFAR-10. Cityscapes contains 2975 training, 500 validation and 1525 testing images with resolution 1024×2048. We present half-resolution experiments which use bilinear interpolation for images and nearest neigbour subsampling for all labels. CIFAR-10 consists of 50000 training and 10000 test images of resolution 32 × 32.

We apply the consistency loss to all training images (including the labeled ones). We train on batches of B_l labeled and B_u unlabeled images. We perform $\lfloor |\mathbb{D}_\mathrm{l}| / B_\mathrm{l} \rfloor$ training steps per epoch without early stopping. We use the same perturbation model across all datasets and tasks, which is likely suboptimal [34].

We train our segmentation models on random 448×448 crops with random scaling and horizontal flipping. The scaling factors are sampled from $\mathcal{U}(1.5^{-1}, 1.5)$. We use $(B_\mathrm{l}, B_\mathrm{u}) = (8, 8)$ for SwiftNet-RN18 [22] and $(B_\mathrm{l}, B_\mathrm{u}) = (4, 4)$ for DeepLab v2 [19]. We train all models for 74400 iterations, which corresponds to 200 epochs with SwiftNet and 100 epochs with DeepLab v2 when all Cityscapes training labels are used. In comparison with SwiftNet-RN18, DeepLab v2 incurs a 12-fold slow-down of per-image throughput during supervised training. However, it also requires less epochs since it has very few parameters with random initialization. Hence, semi-supervised DeepLab v2 trains in 30h on RTX 2080 Ti, which is more than 5 times slower than SwiftNet-RN18. We initialize backbone parameters with public parameterizations pre-trained on ImageNet. To reduce an observed generalization drop when there is a longer period of low learning rates at the end of training, we schedule the learning rate according to $e \mapsto \eta \cos(e\pi/2)$, where $e \in [0..1]$ is the epoch index divided by the total number of epochs. We use $\eta = 4 \cdot 10^{-4}$ for randomly initialized and $\eta = 10^{-4}$ for pre-trained parameters. We use Adam with $(\beta_1, \beta_2) = (0.9, 0.999)$. The L^2 regularization weight is 10^{-4} for randomly initialized and $2.5 \cdot 10^{-5}$ for pre-trained parameters.

We perform classification experiments on CIFAR-10 on WRN-28-2 with standard hyperparameters [35]. We augment all training images with standard random flips and translations. We use $(B_\mathrm{l}, B_\mathrm{u}) = (128, 640)$. We train for 1000 epochs with $|\mathbb{D}_\mathrm{l}| = 4000$ in semi-supervised, and 100 epochs in supervised training.

4.2 Semantic segmentation on Cityscapes

Table 1 compares our models with the state of the art on half-resolution Cityscapes val. The top section presents the previous work [5, 4, 27, 6]. The middle section presents our experiments based on DeepLab v2 [19]. All these experiments use the same splits in order to ensure fair comparison. The first row shows that our experiment with the public code [6] reproduces their accuracy. The last three rows show experiments with our code. We use more training iterations than previous work since that would be a method of choice in all practical applications. Hence, our performance is consistently greater than in the first section of the table. We enable fair comparison with [6] by plugging their method into our training procedure. Under these conditions, our MT-PhTPS outperforms MT-CutMix with L2 loss and confidence thresholding for 1.8 to 2.7 percentage points (pp) with 1/4 to 1/1 of labels, while undeperforming within variance with 1/8 of labels.

The bottom section presents experiments based on SwiftNet-RN18 as mean and standard deviations across 5 different subsets. The last two rows again show that our perturbation model outperforms CutMix when 1/4 or more labels are available. Notably, our perturbation model succeeds to improve upon the fully supervised baseline for both backbones. We observe that DeepLab v2 gets more benefit from unsupervised loss and comes out slightly better in most semi-supervised experiments, in spite of being worse in the supervised setup. This makes sense due to a stronger backbone (ResNet-101 vs ResNet-18) and much more capacity. Nevertheless, SwiftNet-RN18 comes out as a clear method of choice for applications due to 12-fold faster inference. We also note that Mean Teacher performs comparably to simple consistency in experiments with all labeled data.

4.3 Validation of consistency variants

Table 2 compares supervised baselines with 4 kinds of unsupervised consistency on CIFAR-10 and half-resolution Cityscapes. We investigate the following

Method	Proportion of labels used 1/8	 1/4	 1/2	 1/1
DL supervised [5, 6]	56.2	60.2	–	66.0
DL s4GAN [5]	59.3	61.9		65.8
DL AdvSemSeg [4]	58.8	62.3	65.7	67.7
DL ECS [27]	60.3	63.8	–	67.7
DL MT-CutMix [6]	$60.3_{1.2}$	$63.9_{0.7}$	–	$67.7_{0.4}$
DL supervised$_{[6]}$	54.7	59.7	64.6	67.5
DL supervised	56.7	62.5	67.8	69.0
DL MT-CutMix$_{\sim[6]}$	62.4	65.0	67.6	69.0
DL MT-PhTPS	61.9	66.8	69.9	71.7
SN supervised	$55.1_{0.9}$	$61.5_{0.5}$	$66.9_{0.7}$	$70.5_{0.6}$
SN simple-CutMix	$59.8_{0.5}$	$63.8_{1.2}$	$67.0_{0.4}$	$69.3_{1.1}$
SN simple-PhTPS	$62.7_{3.5}$	$65.3_{1.9}$	$68.5_{0.6}$	$71.4_{0.6}$
SN MT-CutMix	$59.3_{1.3}$	$63.3_{1.0}$	$66.8_{0.6}$	$69.7_{0.5}$
SN MT-CutMix$_{\sim[6]}$	$61.6_{0.9}$	$64.6_{0.5}$	$67.6_{0.7}$	$69.9_{0.6}$
SN MT-PhTPS	$62.0_{1.3}$	$66.0_{1.0}$	$69.1_{0.5}$	$71.2_{0.7}$

Table 1. Semi-supervised semantic segmentation accuracy (mIoU/%) on half-resolution Cityscapes val with different proportions of labeled data. The top section reviews experiments from previous work. The middle section presents our experiments on a single dataset split with DeepLab v2 (DL). The first row uses code from [6], while other rows use our code. The bottom section presents experiments with our code on SwiftNet-RN18 (SN) and different consistency variants. Here we run experiments on 3 random dataset splits. The subscript "$\sim$[6]" denotes training with L^2 loss, confidence thresholding and $\alpha = 1$ as proposed in [6] instead of KL divergence with $\alpha = 0.5$.

kinds of consistency: one-way with perturbed teacher input (1w-pt), one-way with perturbed student input (1w-ps), two-way with one perturbed input (2w-p1), and one-way with both inputs perturbed (1w-p2). Note that two-way consistency is not possible with Mean Teacher. All experiments use PhTPS perturbations. CIFAR-10 experiments train on 4000 labels and 50000 images. Cityscapes experiments correspond to the setup from Table 1 based on SwiftNet-RN18. The table shows that 1w-ps performs best, while 2w-p1 performs in-between 1w-ps and 1w-pt. This supports the hypothesis from 3.2, that predictions from unperturbed inputs represent better targets for our unsupervised loss. The 1w-p2 setup underperforms with respect to the baseline, but often outperforms 1w-pt. A closer inspection reveals that 1w-p2 sometimes learns to cheat the consistency loss by outputting similar predictions in all perturbed images. This seems to occur more often when batch normalization uses batch statistics. However, we do not observe the cheating with CutMix. The worst performer is 1w-pt with simple consistency. A closer inspection of Cityscapes experiments reveals severe overfitting to the training dataset as well as consistency cheating.

Configuration	sup.	1w-ps	1w-pt	2w-p1	1w-p2
simple-C10-4k	$80.8_{0.4}$	$90.8_{0.3}$	$50.1_{20.1}$	$73.3_{7.0}$	$72.9_{1.0}$
MT-C10-4k	$80.8_{0.4}$	$90.8_{0.4}$	$80.5_{0.5}$	-	$73.4_{1.4}$
simple-CS-1/4	$61.5_{0.5}$	$65.3_{1.9}$	$1.6_{1.0}$	$16.7_{3.0}$	$61.8_{0.8}$
MT-CS-1/4	$61.5_{0.5}$	$66.0_{1.0}$	$61.5_{1.4}$	-	$61.8_{1.0}$

Table 2. Comparison of 4 consistency variants under PhTPS perturbations: (1) one-way with perturbed teacher input (1w-pt), (2) one-way with perturbed student input (1w-ps), (3) two-way with one input perturbed (2w-p1), and (4) one-way with both inputs perturbed (1w-p2). Algorithms are evaluated on CIFAR-10 test (accuracy/%) while training on 4000 out of 50000 labels (C10-4k) and half-resolution Cityscapes val (mIoU/%) while training on 1/4 of labels from Cityscapes train with SwiftNet-RN18 (CS-1/4).

Semi-supervised experiments on the CamVid dataset [36] resulted in similar relations between consistency variants from Table 2. However, a weaker unsupervised loss with weaker perturbations was required for improving upon the supervised baseline.

5 Conclusion

We have presented a method for semi-supervised semantic segmentation which achieves competitive accuracy in combination with two convolutional models. We show that one-way consistency with unperturbed teacher has two important advantages: i) it has the same training footprint as the standard supervised setup, and ii) it results in better generalization due to learning with less noise. Experiments with many labeled images indicate that simple one-way consistency may outperform Mean Teacher.

To the best of our knowledge, this is the first account of semi-supervised semantic segmentation with efficient models. This combination is essential for many practical real-time applications where there is a lack of large datasets with suitable pixel-level groundtruth.

Suitable directions for future work include further research in semi-supervised learning of efficient dense prediction models. Mild memory requirements will especially favor derivative works for semi-supervised dense prediction in video.

Acknowledgements

This work has been supported by Croatian Science Foundation, grant IP-2020-02-5851 ADEPT. This work has also been supported by European Regional Development Fund, grant KK.01.2.1.02.0119 DATACROSS, and by VSITE College for Information Technologies who provided access to 1 GPU Tesla-V100 32GB.

References

[1] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[2] Antti Tarvainen and Harri 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.

[3] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5689–5697. IEEE Computer Society, 2017.

[4] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. In *BMVC*, page 65, 2018.

[5] S. Mittal, M. Tatarchenko, and T. Brox. Semi-supervised semantic segmentation with high- and low-level consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019.

[6] Geoffrey French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *BMVC*, 2020.

[7] Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen de Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *MICCAI*, 2019.

[8] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):1979–1993, 2019.

[9] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc 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.

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset. In *CVPRW*, 2015.

[11] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. Mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, pages 5000–5009, 2017.

[12] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2017, Fort Worth, TX, USA, July 23-28, 2017*, pages 3226–3229. IEEE, 2017.

[13] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in neural information processing systems*, pages 766–774, 2014.

[14] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2226–2234, 2016.

[15] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7472–7481. IEEE Computer Society, 2018.

[16] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*, pages 135–152, 2018.

[17] Jizong Peng, Guillermo Estrada, Marco Pedersoli, and Christian Desrosiers. Deep co-training for semi-supervised image segmentation. *Pattern Recognit.*, 107:107269, 2020.

[18] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. In-place activated batchnorm for memory-optimized training of DNNs. In *CVPR*, June 2018.

[19] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018.

[20] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and CV Jawahar. Universal semi-supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5259–5270, 2019.

[21] Mengshi Qi, Yunhong Wang, Jie Qin, and Annan Li. Ke-gan: Knowledge embedded generative adversarial networks for semi-supervised scene parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5237–5246, 2019.

[22] Marin Orsic, Ivan Kreso, Petra Bevandic, and Sinisa Segvic. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 12607–12616, 2019.

[23] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. 63(12), 2020.

[24] Dong hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICMLW*, 2013.

[25] Yi Zhu, Zhongyue Zhang, Chongruo Wu, Zhi Zhang, Tong He, Hang Zhang, R. Manmatha, Mu Li, and Alexander J. Smola. Improving semantic segmentation via self-training. *CoRR*, abs/2004.14960, 2020.

[26] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D. Collins, Ekin D. Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Leveraging semi-supervised learning in video sequences for urban scene segmentation. *CoRR*, abs/2005.10266, 2020.

[27] Robert Mendel, Luis Souza Jr, David Rauber, João Papa, and Christoph Palm. Semi-supervised segmentation based on error-correcting supervision. In *ECCV*, 2020.

[28] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020.

[29] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *ECCV*, volume 11207, pages 418–434, 2018.

[30] Ping Chao, Chao-Yang Kao, Yu-Shan Ruan, Chien-Hsiang Huang, and Youn-Long Lin. Hardnet: A low memory traffic network. In *ICCV*, pages 3551–3560, 2019.

[31] Dong Nie, Jia Xue, and Xiaofeng Ren. Bidirectional pyramid networks for semantic segmentation. In *ACCV*, volume 12622 of *Lecture Notes in Computer Science*, pages 654–671. Springer, 2020.

[32] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(6):567–585, 1989.

[33] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[34] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. *arXiv preprint arXiv:1909.13719*, 2019.

[35] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016.

[36] Gabriel J Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.