Curriculum Semi-supervised Segmentation

Hoel Kervadec[(✉)], Jose Dolz, Éric Granger, and Ismail Ben Ayed

ÉTS Montréal, Montréal, Canada
`hoel.kervadec.1@estmtl.net`

Abstract. This study investigates a curriculum-style strategy for semi-supervised CNN segmentation, which devises a regression network to learn image-level information such as the size of the target region. These regressions are used to effectively regularize the segmentation network, constraining the softmax predictions of the unlabeled images to match the inferred label distributions. Our framework is based on inequality constraints, which tolerate uncertainties in the inferred knowledge, e.g., regressed region size. It can be used for a large variety of region attributes. We evaluated our approach for left ventricle segmentation in magnetic resonance images (MRI), and compared it to standard proposal-based semi-supervision strategies. Our method achieves competitive results, leveraging unlabeled data in a more efficient manner and approaching full-supervision performance.

Keywords: Image segmentation · Semi-supervised learning · Constrained CNNs

1 Introduction

In the recent years, deep learning architectures, and particularly convolutional neural networks (CNNs), have achieved state-of-the-art performances in a breadth of visual recognition tasks. These architectures currently dominate the literature in medical image segmentation [12]. The generalization capabilities of these networks typically rely on large and annotated datasets, which, in the case of segmentation, consist of precise pixel-level annotations. Obtaining expert annotations in medical images is a costly process that also requires clinical expertise. The lack of large annotated datasets has driven research in deep segmentation models that rely on reduced supervision for training, such as weakly [8,9,11,17] or semi-supervised [1,19] learning. These strategies assume that annotations are limited or coarse, such as image-level tags [15,17], scribbles [20] or bounding-boxes [18].

In this paper, we focus on semi-supervised learning, a common scenario in medical imaging, where a small set of images are assumed to be fully annotated, but an abundance of unlabeled images is available. Recent progress of these techniques in medical image segmentation has been bolstered by deep learning [1,2,6,14,19,24]. Self-training is a common semi-supervised learning strategy, which consists of employing reliable predictions generated by a deep learning architecture to re-train

it, thereby augmenting the training set with these predictions as pseudo-labels [1,17,18]. Although this approach can leverage unlabeled images, one of its main drawbacks is that early mistakes are propagated back to the network, being re-amplified during training [4,25]. Several techniques were proposed to overcome this issue, such as co-training [24] and adversarial learning [5,13,23]. Nevertheless, with these approaches, training typically involves several networks, or multiple objective functions, which might hamper the convergence of such models.

Alternatively, some weakly supervised segmentation approaches have been proposed to constrain the network predictions with global label statistics, for example, in the form of target-region size [7,8,17]. For instance, Jia et *al.* [7] employed an $\mathcal{L}_2$ penalty to impose equality constraints on the size of the target regions in the context of histopathology image segmentation. However, their formulation requires the exact knowledge of region size, which limits its applicability. More recently, Kervadec et *al.* [8] proposed using inequality constraints, which provide more flexibility, and significantly improves performance compared to cases where learning relies on partial image labels in the form of scribbles. Nevertheless, the values used to bound network predictions in [8] are derived from manual annotations, which is a limiting assumption. Another closely related work is the curriculum learning strategy proposed in the context of unsupervised domain adaptation for urban images in [22]. In this case, the authors proposed to match global label distributions over source (*labelled*) and target (*unlabelled*) images by minimizing the KL-divergence between distributions. Finally, it is worth noting that the semi-supervised learning technique in [6] embeds semantic constraints on the adjacency graph of a given region.

Inspired by this research, we propose a curriculum-style strategy for deep semi-supervised segmentation, which employs a regression network to predict image-level information such as the size of the target region. These regressions are used to effectively regularize the segmentation network, enforcing the predictions for the unlabeled images to match the inferred label distributions. Contrary to [22], our framework uses inequality constraints, which provides greater flexibility, allowing uncertainty in the inferred knowledge, e.g., regressed region size. Another important difference is that the proposed framework can be used for a large variety of region attributes (e.g., shape moments). We evaluated our approach in the task of left ventricle segmentation in magnetic resonance images (MRI), and compared it to standard proposal-based semi-supervision strategies. Our method achieves very competitive results, leveraging unlabeled data in a more efficient manner and approaching full-supervision performance. We made our code publicly available[1].

2   Self-training for Semi-supervised Segmentation

Let $X : \Omega \subset \mathbb{R}^{2,3} \rightarrow \mathbb{R}$ denotes a training image, with Ω its spatial domain. Consider a semi-supervised scenario with two subsets: $\mathcal{S} = \{(X_i, Y_i)\}_{i=1,...,n}$

---

[1] https://github.com/LIVIAETS/semi_curriculum.

which contains a set of images X_i and their corresponding pixel-wise ground-truth labels Y_i, and $\mathcal{U} = \{X_j\}_{j=1,\dots,m}$ a set of unlabeled images, with $m \gg n$. In the fully supervised setting, training is formulated as minimizing the following loss with respect to network parameters $\boldsymbol{\theta}$:

$$\mathcal{L}_Y(\boldsymbol{\theta}) = -\sum_{i \in \mathcal{S}} \sum_{p \in \Omega} Y_{i,p} \log S(X_i|\boldsymbol{\theta})_p \qquad (1)$$

where $S(X_i|\boldsymbol{\theta})_p$ represents a vector of softmax probabilities generated by the CNN at each pixel p and image i. To simplify the presentation, we consider the two-region segmentation scenario (i.e., two classes), with ground-truth binary labels $Y_{i,p}$ taking values in $\{0, 1\}$, 1 indicating the target region (foreground) and 0 indicating the background. However, our formulation can be easily extended to the multi-region case. Common approaches for semi-supervised segmentation [1, 15] generate fake full masks (segmentation proposals) $\tilde{Y}$ for the unlabeled images, which are then used iteratively for network training by adding a standard cross-entropy loss of the form in Eq. (1): $\min_{\boldsymbol{\theta}} \mathcal{L}_Y(\boldsymbol{\theta}) + \mathcal{L}_{\tilde{Y}}(\boldsymbol{\theta})$. The process consists of alternating segmentation-proposal generation and updating network parameters using both labeled data and the new generated masks. Typically such proposals are refined with additional priors such as dense CRF [20]. However, errors in such proposals may mislead training as the cross-entropy loss is minimized over mislabled points and, reinforcing early mistakes during training, as is well-known in the semi-supervised learning literature [4, 25].

3   Curriculum Semi-supervised Learning

The general principle of curriculum learning consists of solving easy tasks first in order to infer some necessary properties about the unlabeled images. In particular, the first task is to learn image-level properties, e.g. the size of the target region, which is easier than learning pixelwise segmentations within an exponentially large label space. Then, we use such image-level properties to facilitate segmentation via constrained CNNs. Figure 1 depicts an illustration of our curriculum semi-supervised segmentation. We first use an auxiliary network that predicts the target-region size for a given image. Particularly, we train a regression network R (with parameters $\tilde{\boldsymbol{\theta}}$) by solving the following minimization problem:

$$\min_{\tilde{\boldsymbol{\theta}}} \sum_{i \in \mathcal{S}} \left(R(X_i|\tilde{\boldsymbol{\theta}}) - \sum_{p \in \Omega} Y_{i,p} \right)^2. \qquad (2)$$

This amounts to minimizing the squared difference between the predicted size and the actual region size.

Now we can define our constrained-CNN segmentation problem using auxiliary size predictions $R(X_i|\tilde{\boldsymbol{\theta}})$:

$$\min_{\boldsymbol{\theta}} \mathcal{L}_Y(\boldsymbol{\theta})$$

$$\text{s.t.} \quad \forall i \in \mathcal{U} : (1-\gamma)R(X_i|\tilde{\boldsymbol{\theta}}) \leq \sum_{p \in \Omega} S(X_i|\boldsymbol{\theta})_p \leq (1+\gamma)R(X_i|\tilde{\boldsymbol{\theta}}), \qquad (3)$$

where the inequality constraints impose the learned image-level information (i.e., region size) on the outputs of the segmentation network for unlabeled images, and γ is a hyper-parameter controlling constraints tightness. We use a penalty-based approach [8] for handling the inequality constraints, which accommodates standard stochastic gradient descent. This amounts to replacing the constraints in (3) with the following penalty over unlabeled samples:

$$\mathcal{L}_{\mathcal{U}}(\boldsymbol{\theta}) = \sum_{i \in \mathcal{U}} \mathcal{C} \left(\sum_{p \in \Omega} S(X_i|\boldsymbol{\theta})_p \right) \tag{4}$$

$$\mathcal{C}(t) = \begin{cases} (t - (1-\gamma)R(X_i|\tilde{\boldsymbol{\theta}}))^2 & \text{if } t \leq (1-\gamma)R(X_i|\tilde{\boldsymbol{\theta}}) \\ (t - (1+\gamma)R(X_i|\tilde{\boldsymbol{\theta}}))^2 & \text{if } t \geq (1+\gamma)R(X_i|\tilde{\boldsymbol{\theta}}) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

This gives our final unconstrained optimization problem: $\min_{\boldsymbol{\theta}} \mathcal{L}_Y(\boldsymbol{\theta}) + \lambda \mathcal{L}_{\mathcal{U}}(\boldsymbol{\theta})$, with λ a hyper-parameter controlling the relative contribution of each term.
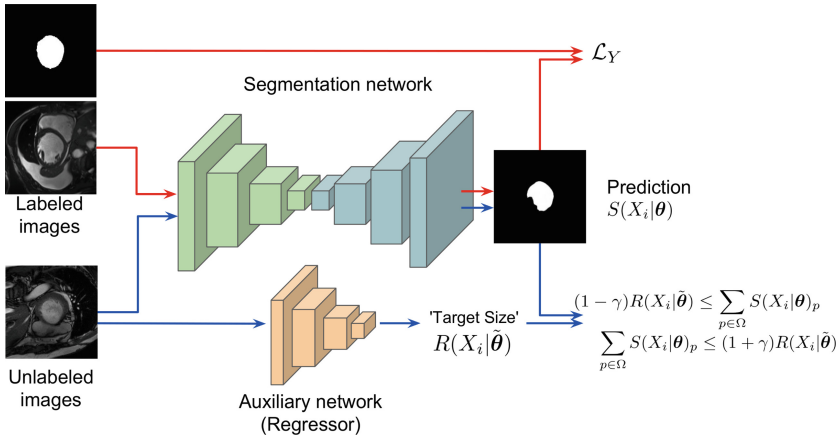

Fig. 1. Illustration of our curriculum semi-supervised segmentation strategy.

4 Experiments

4.1 Setup

Data. Our experiments focused on left ventricular endocardium segmentation. We used the training set from the publicly available data of the 2017 ACDC Challenge [3]. This set consists of 100 cine magnetic resonance (MR) exams covering well defined pathologies: dilated cardiomyopathy, hypertrophic cardiomyopathy, myocardial infarction with altered left ventricular ejection fraction and

abnormal right ventricle. It also included normal subjects. Each exam only contains acquisitions at the diastolic and systolic phases. We sliced and resized the exams into 256×256 images. No additional pre-processing was performed.

Training. For the experiments, we employed 75 exams for training and the remaining 25 for validation. From the training set, we consider that n images are fully annotated and the pixel-wise annotations of the remaining 75-n images are unknown. The n images, and their corresponding ground truth, are employed to train both the auxiliary size predictor and the main segmentation network, in a separate way. To validate both networks, we split the validation set into two smaller subsets of 5 and 20 exams, respectively. The training set undergoes data augmentation only to train the size regressor, by flipping, mirroring and rotating (up to $45°$) the original images, obtaining a training set that is 10 times larger.

Implementation Details. We employed ResNeXt 101 [21] as the backbone architecture for our regressor model, with the squared $\mathcal{L}_2$ norm as the objective function. We trained via standard stochastic gradient descent, with a learning rate of 5×10^{-6}, a momentum of 0.9 and a weight decay of 10^{-4}, for 200 epochs. The learning rate was halved at epochs 100 and 150. We used a batch size of 10. We used ENet [16] as the segmentation network, trained with Adam [10], a learning rate of 5×10^{-4}, $\beta_1 = 0.9$ and $\beta_2 = 0.99$ for 100 epochs. The learning rate was halved if validation DSC did not improve for 20 epochs. We used a batch size of 1, and γ from Eq. (4) is set at $\gamma = 0.1$. We did not use any form of post-processing on the network output.

Comparative Methods. We compare the performance of the proposed semi-supervised curriculum segmentation approach to several models. First, we train a network using only n exams and their corresponding pixel-wise annotations, which is referred to as *FS*. Then, once this model is trained, and following standard proposal-based strategies for semi-supervision, e.g., [1], we perform the inference on the remaining 75-n exams, and include the CNN predictions in the training set, which serve as pseudo-labels for the non-annotated images (referred to as *Proposals*). In this particular case, the training reduces to minimizing the cross-entropy over all the pixels in the manually annotated images and over the pixels predicted as left-ventricle in the pseudo-labels. Since we investigate how to leverage unlabeled data only by learning from the subset of labeled data, we do not integrate any additional cues during training, such as Conditional Random Fields (CRF)[2]. Finally, we train a model with the exact size derived from the ground truth for each image, as in [8], which will serve as an upper bound, referred to as *Oracle*.

Evaluation. We resort to the common dice (DSC) overlap metric between the ground truth and the CNN segmentation to evaluate the performances of the segmentation models. More specifically, we report the mean and standard deviation of the validation DSC over the last 50 epochs of training.

---

[2] Note that the proposal-based methods in [1] use CRF to boost performance.

4.2   Results

We report in Table 1 and Fig. 2 the quantitative evaluation of the different segmentation models. First, we can observe that integrating the size predicted on unlabeled images by the auxiliary network improves the performance compared to solely training from labeled images. The gap is particularly significant when few annotated images are available, ranging from nearly 15 to 25% of difference in terms of DSC. As more labeled images are available, the proposed strategy still improves the performance of the fully supervised counterpart, but by a smaller margin, which goes from 1 to 3%. Compared to the *Oracle*, our method achieves comparable results as the number of training samples increases. This suggests that, when few annotated patients are available, having a better estimation of the size helps to better regularize the network. It is noteworthy to mention that in the *Oracle*, the exact size is known for each image, which results in extra supervision compared to the proposed method. The *proposals* method achieves the same or worse results than its *FS* counterpart, for all the n values evaluated. These results indicate that n patients are not sufficient to train an auxiliary network that generates usable pseudo-labels, due to the difficulty of the segmentation task. This confirms that training a network on an easier task, e.g., learning the size of the target region, can guide the training in a semi-supervised setting.

Table 1. Quantitative results for the different models. Values represent the mean Dice (and standard deviation) over the last 50 epochs.

# labeled patients	Method			
	FS	Proposals	Proposed	Oracle [8]
5	24.8 (4.9)	8.1 (0.8)	53.1 (3.0)	74.3 (2.5)
10	44.4 (8.3)	43.9 (2.9)	58.5 (3.6)	75.7 (3.9)
20	71.7 (3.2)	49.1 (5.0)	72.7 (1.6)	79.0 (2.5)
30	73.1 (1.7)	62.6 (4.4)	75.4 (1.6)	77.0 (1.9)
40	75.8 (2.4)	68.8 (5.6)	76.3 (2.1)	80.4 (2.1)
75	81.6 (1.9)	NA	NA	NA

Evolution of DSC on the validation set over training for some models is depicted in Fig. 3. From these plots, we can observe that the auxiliary network facilitates the training of a harder task, consistently achieving higher performance and better stability than its *FS* counterpart, especially when few labeled images are available. Regarding the instability of the *FS* method, it may be caused by the small number of samples employed for training, with no other source of information that regularizes the network.
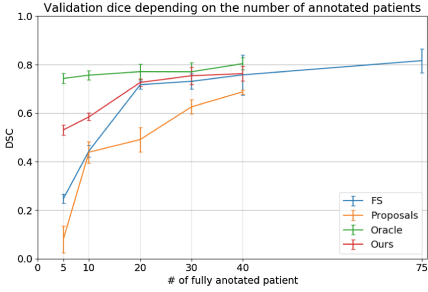
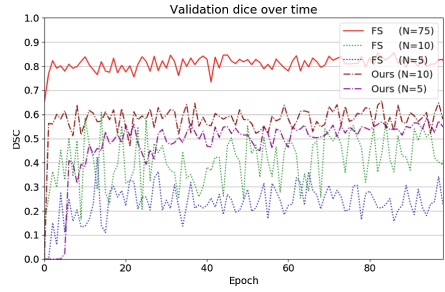**Fig. 2.** Mean DSC per method and for several n annotated patients.

Fig. 3. Validation DSC over time, with a subset of the evaluated models.

Qualitative results are depicted in Fig. 4. Particularly, we show the prediction on the same slice with the different methods and for increasing n. We first observe that predictions of the *FS* model are very unstable, not clearly improving as more labeled images are included in the training, which aligns with the results found in Fig. 3. Then, the *Proposals* approach fails to generate visually acceptable segmentations, even with 30 pixel-wise labeled patients. Although
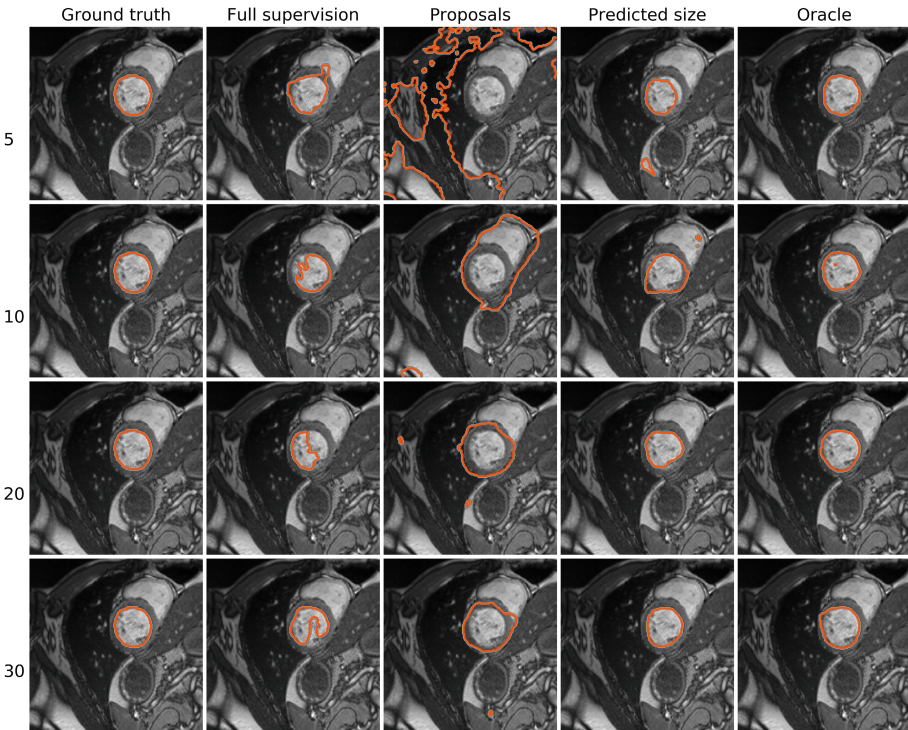

Fig. 4. Visual comparison for the different methods, with varying number of fully annotated patients used for training. Best viewed in colors

its performance improves with the number of labeled patients used in training, its results are not visually satisfying for any value of n. Our curriculum semi-supervised segmentation approach achieves decent results from $n = 5$. It only requires 20 patients to yield comparable segmentations to those of the *Oracle* and the manual ground truth.

References

1. Bai, W., et al.: Semi-supervised learning for network-based cardiac MR image segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 253–260. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_29

2. Baur, C., Albarqouni, S., Navab, N.: Semi-supervised deep learning for fully convolutional networks. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 311–319. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_36

3. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE TMI **37**(11), 2514–2525 (2018)

4. Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised learning. IEEE Trans. Neural Netw. **20**(3), 542–542 (2009). (Chapelle, O. et al. (eds.) 2006) [book reviews]

5. Dong, N., Kampffmeyer, M., Liang, X., Wang, Z., Dai, W., Xing, E.: Unsupervised domain adaptation for automatic estimation of cardiothoracic ratio. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 544–552. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_61

6. Ganaye, P.-A., Sdika, M., Benoit-Cattin, H.: Semi-supervised learning for segmentation under semantic constraint. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11072, pp. 595–602. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00931-1_68

7. Jia, Z., Huang, X., Chang, E.I., Xu, Y.: Constrained deep weak supervision for histopathology image segmentation. IEEE TMI **36**(11), 2376–2388 (2017)

8. Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y., Ben Ayed, I.: Constrained-CNN losses for weakly supervised segmentation. Med. Image Anal. **54**, 88–99 (2019)

9. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: weakly supervised instance and semantic segmentation. In: CVPR, pp. 876–885 (2017)

10. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

11. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: ScribbleSup: scribble-supervised convolutional networks for semantic segmentation. In: CVPR, pp. 3159–3167 (2016)

12. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)

13. Mondal, A.K., Dolz, J., Desrosiers, C.: Few-shot 3D multi-modal medical image segmentation using generative adversarial learning. arXiv:1810.12241 (2018)

14. Nie, D., Gao, Y., Wang, L., Shen, D.: ASDNet: attention based semi-supervised deep networks for medical image segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 370–378. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_43

15. Papandreou, G., Chen, L.C., Murphy, K., Yuille, A.L.: Weakly-and semi-supervised learning of a DCNN for semantic image segmentation. In: ICCV (2015)
16. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: ENet: a deep neural network architecture for real-time semantic segmentation. arXiv preprint: arXiv:1606.02147 (2016)
17. Pathak, D., Krahenbuhl, P., Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation. In: ICCV, pp. 1796–1804 (2015)
18. Rajchl, M., et al.: DeepCut: object segmentation from bounding box annotations using convolutional neural networks. IEEE TMI **36**(2), 674–683 (2017)
19. Sedai, S., Mahapatra, D., Hewavitharanage, S., Maetschke, S., Garnavi, R.: Semi-supervised segmentation of optic cup in retinal fundus images using variational autoencoder. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 75–82. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_9
20. Tang, M., Perazzi, F., Djelouah, A., Ben Ayed, I., et al.: On regularized losses for weakly-supervised CNN segmentation. In: ECCV, pp. 507–522 (2018)
21. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: CVPR, pp. 1492–1500 (2017)
22. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: ICCV, pp. 2020–2030 (2017)
23. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 408–416. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_47
24. Zhou, Y., et al.: Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training. In: IEEE WACV, pp. 121–140 (2019)
25. Zhu, X., Goldberg, A.B.: Introduction to semi-supervised learning. Synth. Lect. Artif. Intell. Mach. Learn. **3**(1), 1–130 (2009)