Source-Relaxed Domain Adaptation for Image Segmentation

Mathilde Bateson, Hoel Kervadec, Jose Dolz, Hervé Lombaert, Ismail Ben Ayed

ETS Montréal

Abstract. Domain adaptation (DA) has drawn high interests for its capacity to adapt a model trained on labeled source data to perform well on unlabeled or weakly labeled target data from a different domain. Most common DA techniques require the concurrent access to the input images of both the source and target domains. However, in practice, it is common that the source images are not available in the adaptation phase. This is a very frequent DA scenario in medical imaging, for instance, when the source and target images come from different clinical sites. We propose a novel formulation for adapting segmentation networks, which relaxes such a constraint. Our formulation is based on minimizing a label-free entropy loss defined over target-domain data, which we further guide with a domain-invariant prior on the segmentation regions. Many priors can be used, derived from anatomical information. Here, a class-ratio prior is learned via an auxiliary network and integrated in the form of a KullbackLeibler (KL) divergence in our overall loss function. We show the effectiveness of our prior-aware entropy minimization in adapting spine segmentation across different MRI modalities. Our method yields comparable results to several state-of-the-art adaptation techniques, even though is has access to less information, the source images being absent in the adaptation phase. Our straight-forward adaptation strategy only uses one network, contrary to popular adversarial techniques, which cannot perform without the presence of the source images. Our framework can be readily used with various priors and segmentation problems.

1 Introduction

Semantic segmentation, or the pixel-wise annotation of an image, is a key first step in many clinical applications. Since the introduction of deep learning methods, automated methods for segmentation have outstandingly improved in many natural and medical imaging problems [14]. Nonetheless, the pixel-level ground-truth labelling necessary to train these networks is time-consuming, and deep-learning methods tend to under-perform when trained on a dataset with an underlying distribution different from the target images. To circumvent those impediments, methods learning robust networks with less supervision have been popularized in computer vision.

Domain Adaptation (DA) adresses the transferability of a model trained on an annotated source domain to another target domain with no or minimal annotations. The presence of a domain shift, such as those produced by different protocols, vendors, machines in medical imagining, often leads to a big performance

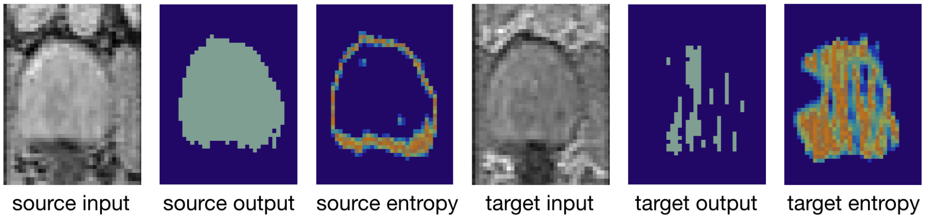

Fig. 1. Visualization of 2 aligned slice pairs in source (Water) and target modality (In-Phase): the domain shift in the target produces a drop in confidence and accuracy.

drop (see Fig. 1). Adversarial strategies are currently the prevailing techniques to adapt networks to a different target domain, both in natural [4,9,10,23] and medical [2,6,7,17] images. These methods can either be generative, by transforming images from one domain to the other [25], or can minimize the discrepancy in the feature or output spaces learnt by the model [3,18,17].

One major limitation of adversarial techniques is that, by design, they require the concurrent access to both the source and target data during the adaptation phase. In medical imaging, this may not be always feasible when the source and target data come from different clinical sites, due to, for instance, privacy concerns or the loss or corruption of source data. Amongst alternative approaches to adversarial techniques, self training [26] and the closely-related entropy minimization [19,20,15] were investigated in computer vision. As confirmed by the low entropy prediction maps in Fig. 1, a model trained on an imaging modality tends to produce very confident predictions on within-sample examples, whereas uncertainty remains high on unseen modalities. As a result, enforcing high confidence in the target domain as well can close the performance gap. This is the underlying motivation for entropy minimization, which was first introduced in semi-supervised [5] and unsupervised [13] learning. To prevent the well-known collapse of entropy minimization to a trivial solution with a single class, the recent domain-adaptation methods in [19,20] further incorporate a criterion encouraging diversity in the prediction distributions. However, similarly to adversarial approaches, the entropy-based methods in [19,20] require access to the source data (both the images and labels) during the adaptation phase via a standard supervised cross-entropy loss. The latter discourages the trivial solution of minimizing the entropy alone on the unlabeled target images.

We propose a domain-adaptation formulation tailored to a setting where the source data is unavailable (neither images, nor labeled masks) during the training of the adaptation phase. Instead, our method only requires the parameters of a model previously trained on the source data for initialisation. Our formulation is based on minimizing an label-free entropy loss defined over target-domain data, which we further guide with a domain-invariant prior on the segmentation regions. Many priors can be used, derived from anatomical information. Here, a class-ratio prior is learned via an auxiliary network and integrated in the form of a KullbackLeibler (KL) divergence in our overall loss function. Unlike

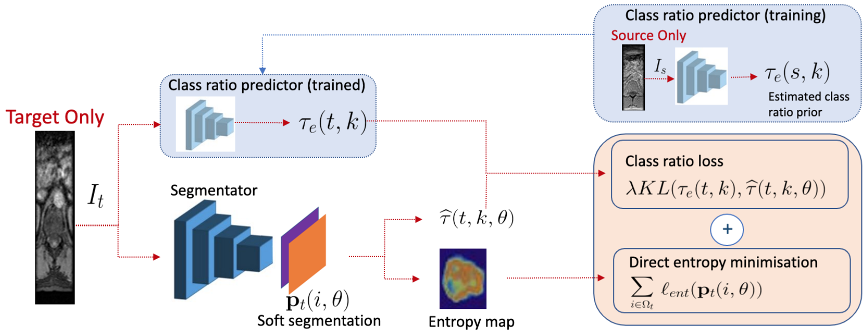
Class ratio predictor (training)
Source Only
I_s
$\tau_e(s,k)$
Estimated class ratio prior
Class ratio predictor (trained)
$\tau_e(t,k)$
Target Only
I_t
Segmentator
$\widehat{\tau}(t,k,\theta)$
$\mathbf{p}_t(i,\theta)$
Soft segmentation
Entropy map
Class ratio loss
$\lambda KL(\tau_e(t,k), \widehat{\tau}(t,k,\theta))$
+
Direct entropy minimisation
$\sum_{i\in\Omega_t} \ell_{ent}(\mathbf{p}_t(i,\theta))$

Fig. 2. Overview of our framework for Source-Relaxed Domain Adaptation: we leverage entropy minimization and a class-ratio prior to relax the need for a concurrent access to the source and target data.

the recent entropy-based methods in [19,20], our overall loss function relaxes the need to access to the source images and labels during adaptation, as we do not use a source-based cross-entropy loss. Our class-ratio prior is related to several recent works in the context of semi- and weakly-supervised learning [24,11], which showed the potential of domain-knowledge priors for guiding deep networks when labeled data is scarce. Also, the recent works in [22,1] integrated priors on class-ratio/size in domain adaptation but, unlike our work, in the easier setting where one has access to source data (both the images and labels). In fact, the works in [22,1] used a cross-entropy loss over labeled source images during the training of the adaptation phase.

We report comprehensive experiments and comparisons with state-of-the-art domain-adaptation methods, which show the effectiveness of our prior-aware entropy minimization in adapting spine segmentation across different MRI modalities. Surprisingly, even though our method does not have access to source data during adaptation, it achieves better performances than the state-of-the-art methods in [22,17], while greatly improving the confidence of network predictions. Our framework can be readily used for adapting any segmentation problems. Our code is publicly and anonymously available [1]. To the best of our knowledge, we are the first to investigate domain adaptation for segmentation without direct access to the source data during the adaptation phase.

2 Method

We consider a set of S source images $I_s : \Omega_s \subset \mathbb{R}^d \rightarrow \mathbb{R}$, $d \in \{2,3\}$, $s = 1, \dots, S$. The ground-truth K-class segmentation of I_s can be written, for each pixel (or voxel) $i \in \Omega_s$, as a simplex vector $\mathbf{y}_s(i) = (y_s^1(i), \dots, y_s^K(i)) \in \{0,1\}^K$. For domain adaptation (DA) problems, the network is usually first trained on the

[1] https://github.com/SRDAMICCAI/SRDA (anonymized)

source domain only, by minimizing a standard supervised loss with respect to network parameters θ:

$$\mathcal{L}_s\left(\theta, \Omega_s\right) = \frac{1}{|\Omega_s|} \sum_{s=1}^{S} \ell\left(\mathbf{y}_s(i), \mathbf{p}_s(i,\theta)\right) \tag{1}$$

where $\mathbf{p}_s(i,\theta) = (p_s^1(i,\theta), \dots, p_s^K(i,\theta)) \in [0,1]^K$ is the softmax output of the network at pixel/voxel i in image I_s, and ℓ is the standard cross-entropy loss: $\ell(\mathbf{y}_s(i), \mathbf{p}_s(i,\theta)) = -\sum_k y_s^k(i) \log p_s^k(i,\theta)$.

Given T images of the target domain, $I_t : \Omega_t \subset \mathbb{R}^{2,3} \rightarrow \mathbb{R}$, $t = 1, \dots, T$, the first loss term in our adaptation phase encourages high confidence in the softmax predictions of the target, which we denote $\mathbf{p}_t(i,\theta) = (p_t^1(i,\theta), \dots, p_t^K(i,\theta)) \in [0,1]^K$. This is done by minimizing the entropy of each of these predictions:

$$\ell_{ent}(\mathbf{p}_t(i,\theta)) = -\sum_k p_t^k(i,\theta) \log p_t^k(i,\theta) \tag{2}$$

However, it is well-known from the semi-supervised and unsupervised learning literature [5,13,8] that minimizing this entropy loss alone may result into degenerate trivial solutions, biasing the prediction towards a single dominant class. To avoid such trivial solutions, the recent domain-adaptation works in [19,20] integrated a standard supervised cross-entropy loss over the source data, i.e., Eq. (1), during the training of the adaptation phase. However, this requires access to the source data (both the images and labels) during the adaptation phase. To relax this requirement, we embed domain-invariant prior knowledge to guide unsupervised entropy training during the adaptation phase, which takes the form of a class-ratio (i.e. region proportion) prior. We express the (unknown) true class-ratio prior for class k and image I_t as: $\tau_{GT}(t,k) = \frac{1}{|\Omega_t|} \sum_{i \in \Omega_t} y_t^k(i)$. As the ground-truth labels are unavailable in the target domain, this prior cannot be computed directly. Instead, we train an auxiliary network on source data to produce an estimation of class-ratio in the target[2], which we denote $\tau_e(t,k)$. Furthermore, the class-ratio can be approximated from the the segmentation network's output for target image I_t as follows: $\widehat{\tau}(t,k,\theta) = \frac{1}{|\Omega_t|} \sum_{i \in \Omega_t} p_t^k(i,\theta)$

To match these two probabilities representing class-ratios, we integrate a KL divergence with the entropy in Eq. (2), minimizing the following overall loss during the training of the adaptation phase:

$$\min_{\theta} \sum_t \sum_{i \in \Omega_t} \ell_{ent}(\mathbf{p}_t(i,\theta)) + \lambda \mathrm{KL}(\tau_e(t,k), \widehat{\tau}(t,k,\theta)) \tag{3}$$

Clearly, minimizing our overall loss in Eq. (3) during adaptation does not use the source images and labels, unlike the recent entropy-based domain adaptation methods in [19,20].

[2] Note that many other estimators could be used, e.g., using region statistics from the source domain or anatomical prior knowledge.

3 Experiments

3.1 Cross-Modality Adaption with entropy minimization

Dataset. We evaluated the proposed method on the publicly available MICCAI 2018 IVDM3Seg Challenge[3] dataset of 16 manually annotated 3D multi-modal magnetic resonance scans of the lower spine, in a study investigating intervertebral discs (IVD) degeneration. We first set the water modality (Wat) as the source and the in-phase (IP) modality as the target domain (Wat → IP), then reverted the adaptation direction (IP → Wat). 13 scans were used for training, and the remaining 3 scans were kept for validation. The slices were rotated in the transverse plane, and no other pre-processing was performed. The setting is binary segmentation (K=2), and the training is done with 2D slices.

Benchmark Methods We compare our loss to the recent one adopted in [22]:

$$\mathcal{L}_s\left(\theta, \Omega_s\right) + \lambda \sum_{t} \sum_{i \in \Omega_T} KL(\tau_e(t,k), \widehat{\tau}_t(t,k,\theta))$$

Note that, in [22], the images from the source and target domain must be present concurrently in this framework, which we denote *AdaSource*. We also compared to the state-of-the art adversarial method in [17], denoted *Adversarial*. A model trained on the source only with Eq(1), *NoAdaptation*, was used as a lower bound. A model trained with the cross-entropy loss on the target domain, referred to as *Oracle*, served as an upper bound.

Learning and estimating the class-ratio prior. We learned an estimation of the ground-truth class-ratio prior via an auxiliary regression network R, which is trained on the images I_s from the source domain S, where the ground-truth class-ratio $\tau_{GT}(s,k)$ is known. R is trained with the squared $\mathcal{L}_2$ loss: $\min_{\tilde{\theta}} \sum_{s=1...S} \left(R(I_s, \tilde{\theta}) - \tau_{GT}(s,k)\right)^2$. The estimated class-ratio prior $\tau_e(t,k)$ of an image I_t in the target domain is obtained by inference. We added weak supervision in the form of image-level tag information by setting $\tau_e(t,k) = (1,0)$ for the target images that do not contain the region of interest. Note that we used exactly the same class-ratio priors and weak supervision for the method in [22], for a fair comparison. Note, also, that we adopted and improved significantly the performance of the adversarial method in [17] by using the same weak supervision information based on image-level tags, for a fair comparison[4].

Training and implementation details. For all methods, the segmentation network employed was ENet [16], trained with the Adam optimizer [12], a batch

[3] https://ivdm3seg.weebly.com/

[4] For the model in [17], pairs of source and target images were not used if neither had the region of interest as this confuses adversarial training, reducing its performance.

size of 12 for 100 epochs, and an initial learning rate of 1×10^{-3}. For all adaptation models, a model trained on the source data with Eq(1) for 100 epochs was used as initialization. The λ parameter in Eq(3) was set empirically to 1×10^{-2}. For *AdaSource*, the batches used were non-aligned random slices in each domain. To learn the class-ratio prior, a ResNeXt101 [21] regression network is used, optimized with SGD, a learning rate of 5×10^{-6}, and a momentum of 0.9.

Evaluation. The Dice similarity coefficient (DSC) and the Hausdorff distance (HD) were used as evaluation metrics in our experiments.

3.2 Discussion

Table 3.2 reports quantitative metrics. As expected, the model *NoAdaptation*, which doesn't use any adaptation strategy but instead is trained using Eq(1) on the source modality, can't perform well on a different target modality. The mean DSC reached is of 46.7% on IP, and 63.7% on Wat, and a very high standard deviation is observed in both case, showing a high subject variability. This is confirmed in Fig. 4, where it can be seen that the output segmentation is poor on 2 out of 3 subjects. Moreover, as can be observed in Fig. 3, where the evolution of the training in terms of validation DSC is shown, the DSC is very unstable on the target domain throughout learning. We also observe that the adaptation task isn't symmetrically difficult, as the performance drop is much bigger in one direction (Wat → IP). The performance of *Oracle*, the upper baseline is also lower in IP. As the visualisation in Fig. 1 show, the higher contrast in Wat images makes the segmentation task easier.

All models using adaptation techniques yield substantial improvement over the lower baseline. First, *Adversarial* achieves a mean DSC of 65.3% on IP and 77.3% on Wat. Nonetheless, the models without adversarial strategies yield better results: *AdaSource* achieves a mean DSC of 67.0% on IP and 78.3% on Wat. Interestingly, our model *AdaEnt* shows comparable performance, with a mean DSC of 67.0% on IP and 77.8% on Wat. These results show that having access to more information (i.e., source data) doesn't necessarily help for the adaptation task. For both models *AdaSource* and *AdaEnt*, the DSC comes close to the *Oracle*'s, the upper baseline, reaching respectively 82% and 82% of its performance respectively on IP, and 87% and 89% of its performance respectively on Wat. This demonstrates the efficiency of the using a class-ratio prior matching with a KL divergence. Moreover, in Fig. 3, we can observe that both these adaptation methods yield rapidly high validation DSC measures (first 20 epochs). This suggests that integrating such a KL divergence helps the learning process in domain adaptation. Finally, the HD values confirm the trend across the different models. Improvement over the lower baseline model (2.45 pixels on IP, 1.44 on Wat), is substantial for both *AdaSource* (1.34 pixels on IP, 1.14 pixels on Wat), as well as for *AdaEnt* (1.33 pixels on IP, 1.17 on Wat).

Qualitative segmentations and corresponding prediction entropy maps are depicted in Figure 4, from the easiest to the hardest subject in the validation

Wat (Source) → IP (Target)		
Method	DSC (%)	HD (pix)
No Adaptation	46.7 ± 10.8	2.45 ± 0.16
Adversarial[17]	65.3 ± 5.5	1.67 ± 1.64
AdaSource [22]	67.0 ± 7.2	1.34 ± 0.15
AdaEnt (Ours)	**67.0** ± **6.1**	**1.33** ± **0.17**
Oracle	82.3 ± 1.2	1.09 ± 0.16

IP (Source) → Wat (Target)		
Method	DSC (%)	HD (pix)
No Adaptation	63.7± 9.1	1.44 ± 0.2
Adversarial[17]	77.3 ± 7.6	1.15 ± 0.2
AdaSource [22]	**78.3** ± **3.5**	**1.14** ± **0.1**
AdaEnt (Ours)	77.8 ± 2.2	1.17 ± 0.1
Oracle	89.0 ± 2.7	0.90 ± 0.1

Table 1. Quantitative comparisons of performance on the *target* domain for the different models (mean ± std) show the efficiency of our source-relaxed formulation.

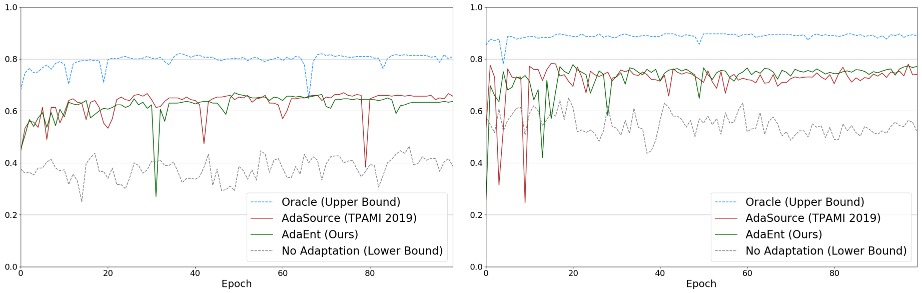

Fig. 3. Evolution of validation DSC over training for the different models. Comparison of the proposed model to the lower and upper bounds, and to the adaptation with access to source in [22] is shown. Water → In-Phase *(left)*, In-Phase → Water *(right)*

set. Without adaptation, a model trained on source data only can't recover the structure of the IVD on the target data, and is very uncertain, as revealed by the high activations in the prediction entropy maps. The output segmentation masks are noisy, with very irregular edges. As expected, the segmentation masks obtained using both adaptation formulations are much closer to the ground truth one, and have much more regular edges. Nonetheless, the entropy maps produced from *AdaSource* predictions still show high entropy activations inside and close to the IVD structures. On the contrary, those produced from *AdaEnt* look like edge detection results with high entropy activations only present along the IVD borders. Interestingly, it can be seen that even the *Oracle*'s segmentation predictions are more uncertain. This isn't surprising, as *AdaEnt* is the only model trained to directly minimize the entropy of the predictions. The visual results confirm *AdaEnt*'s remarkable ability to produce accurate predictions with high confidence.

4 Conclusion

In this paper, we proposed a simple formulation for domain adaptation (DA), which removes the need for a concurrent access to the source and target data, in

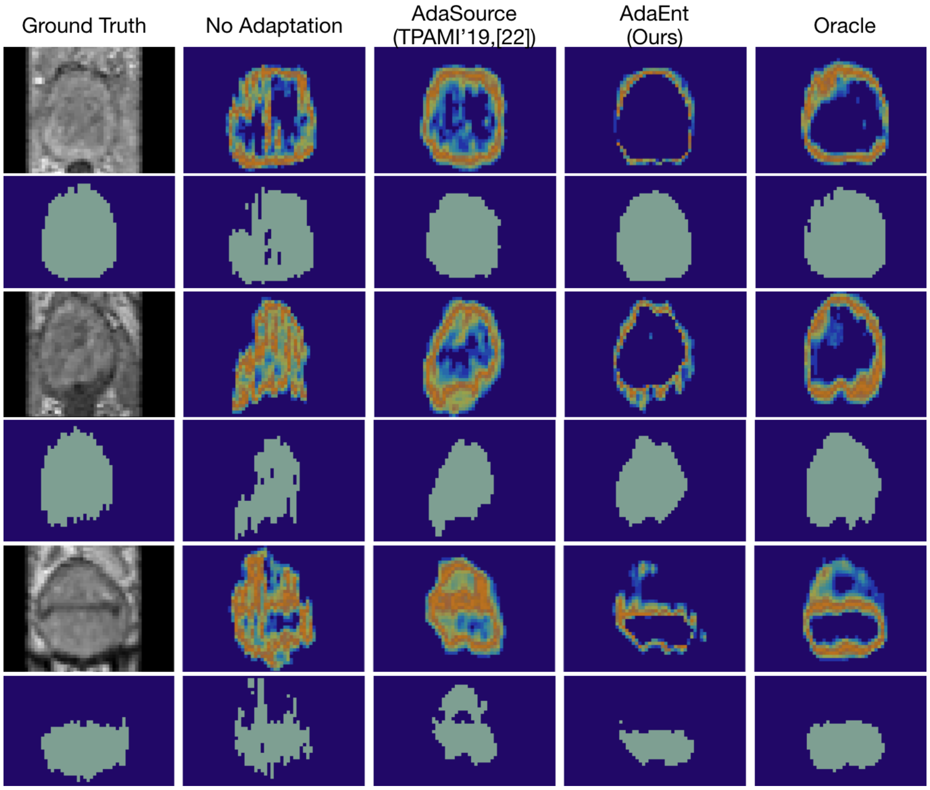

Fig. 4. Visual results for each subject in the validation set for the several models (Water → In-Phase). First column shows an input slice and the corresponding semantic segmentation ground-truth. The other columns show segmentation results (bottom) along with prediction entropy maps produced by the different models (top).

the context of semantic segmentation for multi-modal magnetic resonance images. Our approach substitutes the standard supervised loss in the source domain by a direct minimization of the entropy of predictions in the target domain. To prevent trivial solutions, we integrate the entropy loss with a class-ratio prior, which is built from an auxiliary network. Unlike the recent domain-adaptation techniques, our method tackles DA without resorting to source data during the adaptation phase. Interestingly, our formulation achieved better performances than related state-of-the-art methods with access to both source and target data. This shows the effectiveness of our prior-aware entropy minimization and that, in several cases of interest where the domain shift is not too large, adaptation might not need access to the source data. Our proposed adaptation framework is usable with any segmentation network architecture.

Acknowledgment

This work is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), Discovery Grant program, by the The Fonds de recherche du Qubec - Nature et technologies (FRQNT) grant, the Canada Research Chair on Shape Analysis in Medical Imaging, the ETS Research Chair on Artificial Intelligence in Medical Imaging, and NVIDIA with a donation of a GPU. The authors would like to thank the MICCAI 2018 IVDM3Seg organizers for providing the data.

References

1. Bateson, M., Dolz, J., Kervadec, H., Lombaert, H., Ayed, I.B.: Constrained domain adaptation for segmentation. In: MICCAI (2019)
2. Chen, Y., Li, W., Van Gool, L.: Road: Reality oriented adaptation for semantic segmentation of urban scenes. In: CVPR (2018)
3. Dou, Q., Ouyang, C., Chen, C., Chen, H., Glocker, B., Zhuang, X., Heng, P.: Pnp-adanet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. IEEE Access (2019)
4. Gholami, A., et al.: A novel domain adaptation framework for medical image segmentation. In: MICCAI Brainlesion Workshop (2018)
5. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: NIPS (2004)
6. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: ICML (2018)
7. Hong, W., Wang, Z., Yang, M., Yuan, J.: Conditional generative adversarial network for structured domain adaptation. In: CVPR (2018)
8. Jabi, M., Pedersoli, M., Mitiche, A., Ben Ayed, I.: Deep clustering: On the link between discriminative models and k-means. IEEE TPAMI (2019)
9. Javanmardi, M., Tasdizen, T.: Domain adaptation for biomedical image segmentation using adversarial training. In: ISBI (2018)
10. Kamnitsas, K., et al.: Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: IPMI (2017)
11. Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y., Ayed, I.B.: Constrained-CNN losses for weakly supervised segmentation. MedIA **54** (2019)
12. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. ICLR (2014)
13. Krause, A., Perona, P., Gomes, R.G.: Discriminative clustering by regularized information maximization. In: NIPS (2010)
14. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. MedIA **42** (2017)
15. Morerio, P., Cavazza, J., Murino, V.: Minimal-entropy correlation alignment for unsupervised deep domain adaptation. In: ICLR (2018)
16. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: Enet: A deep neural network architecture for real-time semantic segmentation. arXiv 1606.02147 (2016)
17. Tsai, Y., Hung, W., Schulter, S., Sohn, K., Yang, M., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: CVPR (2018)
18. Tzeng, E., et al.: Adversarial discriminative domain adaptation. In: CVPR (2017)

19. Vu, T.H., Jain, H., Bucher, M., Cord, M., Prez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: CVPR (2019)
20. Wu, X., Zhang, S., Zhou, Q., Yang, Z., Zhao, C., Latecki, L.J.: Entropy minimization vs. diversity maximization for domain adaptation. arXiv 2002.01690 (2020)
21. Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: CVPR (2017)
22. Zhang, Y., David, P., Foroosh, H., Gong, B.: A curriculum domain adaptation approach to the semantic segmentation of urban scenes. IEEE TPAMI (2019)
23. Zhao, H., et al.: Supervised segmentation of un-annotated retinal fundus images by synthesis. IEEE TMI (2019)
24. Zhou, Y., Li, Z., Bai, S., Chen, X., Han, M., Wang, C., Fishman, E., Yuille, A.: Prior-aware neural network for partially-supervised multi-organ segmentation. In: ICCV (2019)
25. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on (2017)
26. Zou, Y., Yu, Z., Kumar, B.V.K.V., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: ECCV (2018)