Constrained Domain Adaptation for Segmentation

Mathilde Bateson(✉), Hoel Kervadec, Jose Dolz, Hervé Lombaert, and Ismail Ben Ayed

ÉTS Montréal, Montreal, Canada
mathilde.bateson.1@ens.etsmtl.ca

Abstract. We propose to adapt segmentation networks with a constrained formulation, which embeds domain-invariant prior knowledge about the segmentation regions. Such knowledge may take the form of simple anatomical information, e.g., structure size or shape, estimated from source samples or known *a priori*. Our method imposes domain-invariant inequality constraints on a network output of unlabeled target samples. It implicitly matches prediction statistics between target and source domains with permitted uncertainty of prior knowledge. We address our constrained problem with a differentiable penalty, fully suited for conventional gradient descent approaches, removing the need for computationally expensive Lagrangian optimization with dual projections. Unlike current two-step adversarial training, our formulation is based on a single loss in a single network, which simplifies adaptation by avoiding extra adversarial steps, while improving convergence and quality of training. The comparison of our approach with state-of-the-art adversarial methods reveals substantially better performance on the challenging task of adapting spine segmentation across different MRI modalities. Our results also show a robustness to imprecision of size priors, approaching the accuracy of a fully supervised model trained directly in a target domain. Our method can be readily used for various constraints and segmentation problems.

Keywords: Image segmentation · Domain adaptation · Constrained CNNs

1 Introduction

Convolutional neural networks (CNNs) are currently dominating segmentation problems, yielding outstanding performances in a breadth of medical imaging applications [14]. A major impediment of such supervised models is that they require large amounts of training data built with scarce expert knowledge and labor-intensive, pixel-level annotations. Typically, segmentation ground truth is available for limited data, and supervised models are seriously challenged with

D. Shen et al. (Eds.): MICCAI 2019, LNCS 11765, pp. 326–334, 2019.
https://doi.org/10.1007/978-3-030-32245-8_37

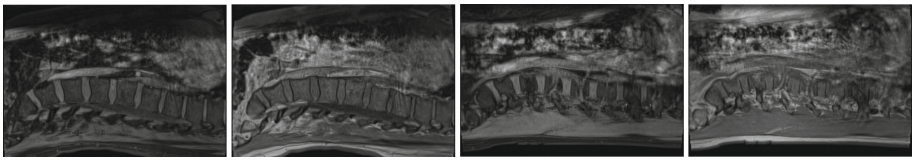

Fig. 1. Visualization of 2 aligned slice pairs in source (Wat) and target modality (IP).

new unlabeled samples (target data) that differ from the labeled training samples (source data) due, for instance, to variations in imaging modalities and protocols, vendors, machines and clinical sites; see Fig. 1. Unsupervised domain adaptation (UDA) tackles such substantial domain shifts between the distributions of the source and target data by learning domain-invariant representations, assuming labels are available only for the source. The subject is currently attracting substantial efforts, both in computer vision [7,20,21] and medical imaging [4,11,18,23]. While a large body of works focused on image classification [19,21], there is a rapidly growing interest into adapting segmentation networks [11,20], more so because building segmentation labels for each new domain is cumbersome.

In the recent literature, adversarial techniques have become the *de facto* choice in adapting segmentation networks, for medical [5,9,11,24] and color [3,7,8,20] images. These techniques match the feature distribution across domains by alternating the training of two networks, one learning a discriminator between source and target features and the other generating segmentations. While adversarial training achieved excellent performances in image classification [21], our experiments suggest that it may not be sufficient for segmentation, where learning a discriminator is much more complex than classification as it involves predictions in an exponentially large label space. This is in line with a few recent works in computer vision [22,25], which argue that adversarial formulations of classification may not be appropriate for segmentation, showing that better performances could be reached via other alternatives, e.g., self training [22] or curriculum learning [22,25]. Furthermore, a large label space might invalidate the assumption that the source and target share the same feature representation at all the abstraction levels of a deep network. In fact, recently, Tsai et al. [20] proposed adversarial training in the softmax-output space, outperforming feature-matching techniques in the context of color images. Such output space conveys domain-invariant information about segmentation structures, for instance, shape and spatial layout, even when the inputs across domains are substantially different. Finally, it is worth mentioning the recent classification study in [19], which argued that adversarial training is not sufficient for high-capacity models, as is the case for segmentation. For deep architectures, the authors of [19] showed experimentally that jointly minimizing source generalization error and feature divergence does not yield high accuracy on the target task.

We propose a general constrained domain adaptation formulation, which embeds domain-invariant prior knowledge about the segmentation regions. Such

knowledge takes the form of simple anatomical information, e.g., region size or shape, which is either estimated from the source ground truth or known *a priori*. For instance, in the application we tackle in our experiments, we can use human-spine measurements that are well known in the literature [1] for constraining the sizes of the inter-vertebral discs in axial MRI slices. By imposing domain-invariant inequality constraints on the network outputs of unlabeled target samples, our method matches implicitly some prediction statistics of the target to the source, and allows uncertainty in the prior knowledge. We address our constrained problem with a differentiable penalty, which can be fully handled with SGD, removing the need for computationally expensive Lagrangian optimization with dual projections. Unlike two-step adversarial training, our method uses a single loss/network, which simplifies adaptation by avoiding extra adversarial steps, while improving training quality and efficiency. We juxtapose our approach to the state-of-art adversarial method in [20] on the challenging task of adapting spine segmentation across different MRI modalities. Our method achieves significantly better performances using simple and imprecise size priors, with a 16% improvement, approaching the performance of a supervised model. It can be readily used for various constraints and segmentation problems. Our code is publicly (and anonymously) available[1].

2 Formulation

Let $I_s : \Omega_s \subset \mathbb{R}^{2,3} \rightarrow \mathbb{R}$, $s = 1, \ldots, S$, denote the training images of the source domain. Assume that each of these has a ground-truth segmentation, which, for each pixel (or voxel) $i \in \Omega_s$, takes the form of binary simplex vector $\mathbf{y}_s(i) = (y_s^1(i), \ldots, y_s^K(i)) \in \{0,1\}^K$, with K the number of classes (segmentation regions).

Given T unlabeled images of the target domain, $I_t : \Omega_t \subset \mathbb{R}^{2,3} \rightarrow \mathbb{R}$, $t = 1, \ldots, T$, we state unsupervised domain adaptation for segmentation as the following constrained optimization w.r.t parameters θ:

$$\begin{aligned} &\min_{\theta} \sum_{s} \sum_{i \in \Omega_s} \mathcal{L}(\mathbf{y}_s(i), \mathbf{p}_s(i,\theta)) \\ &\text{s.t.} \quad f_c(\mathbf{P}_t(\theta)) \leq 0 \quad c = 1, \ldots, C; t = 1, \ldots, T \end{aligned} \tag{1}$$

where $\mathbf{p}_x(i,\theta) = (p_x^1(i,\theta), \ldots, p_x^K(i,\theta)) \in [0,1]^K$ is the softmax output of the network at pixel/voxel i in image $x \in \{t = 1, \ldots, T\} \cup \{t = 1, \ldots, S\}$, and $\mathbf{P}_x(\theta)$ is a $K \times |\Omega_x|$ matrix whose columns are the vectors of network outputs $\mathbf{p}_x(i,\theta), i \in \Omega_x$. In problem (1), $\mathcal{L}$ is a standard loss, e.g., the cross-entropy: $\mathcal{L}(\mathbf{y}_s(i), \mathbf{p}_s(i,\theta)) = -\sum_k y_s^k(i) \log p_s^k(i,\theta)$, computed on the source domain S. The inequality constraint can embed very useful prior knowledge that is invariant across domains and modalities, and is imposed on the network outputs for unlabeled target-domain data. Assume, for instance, that we have prior knowledge about the size (or cardinality) of the target segmentation region (or class) k.

[1] https://github.com/CDAMICCAI2019/CDA.

Such a knowledge is invariant w.r.t modalities, and does not have to be precise; it can be in the form of lower and upper bounds on region size. For instance, when we have an upper bound a on the size of region k, we can impose the following constraint: $\sum_{i\in\Omega_t} p_t^k(i,\theta) - a \leq 0$. In this case, the corresponding constraint c in the general-form constrained problem (1) uses particular function $f_c(\mathbf{P}_t(\theta)) = \sum_{i\in\Omega_t} p_t^k(i,\theta) - a$. In a similar way, one can impose a lower bound b on the size of region k using $f_c(\mathbf{P}_t(\theta)) = b - \sum_{i\in\Omega_t} p_t^k(i,\theta)$. Priors a and b can be learned from the ground-truth segmentations of the source domain (assuming such priors are invariant across domains). Also, depending on the application, such priors may correspond to anatomical knowledge. For instance, in the application we tackle in our experiments, we can use human spine measurements that are well known in the clinical literature [1] for constraining the sizes of the intervertebral discs in axial MRI slices. Our framework can be easily extended to more descriptive constraints, e.g., invariant shape moments [13], which do not change from one modality to another[2].

Even when the constraints are convex with respect to the network probability outputs, the problem in (1) is challenging for deep segmentation models that involve millions of parameters. In the general context of optimization, a standard technique to deal with hard inequality constraints is to solve the Lagrangian primal and dual problems in an alternating scheme [2]. For problem (1), this amounts to alternating the optimization of a CNN for the primal with stochastic optimization, e.g., SGD, and projected gradient-ascent iterates for the dual. However, despite the clear benefits of imposing hard constraints on CNNs, such a standard Lagrangian-dual optimization is avoided in the context of modern deep networks due, in part, to computational-tractability issues. As pointed out in [15, 17], there is a consensus within the community that imposing hard constraints on the outputs of deep CNNs that are common in modern image analysis problems is impractical: The use of Lagrangian-dual optimization for networks with millions of parameters requires training a whole CNN after each iterative dual step.

In the context of deep networks, equality or inequality constraints are typically handled in a "soft" manner by augmenting the loss with a *penalty* function [6,10,12]. The penalty-based approach is a simple alternative to Lagrangian optimization, and is well-known in the general context of constrained optimization; see [2], Sect. 4. In general, such penalty-based methods approximate a constrained minimization problem with an unconstrained one by adding a term, which increases when the constraints are violated. This is convenient for deep networks because it removes the requirement for explicit Lagrangian-dual optimization. The inequality constraints are fully handled within stochastic optimization, as in standard unconstrained losses, avoiding gradient ascent iterates/projections over the dual variables and reducing the computational load for training. For this work, we pursue a similar penalty approach, and replace constrained problem (1) by the following unconstrained problem:

[2] In fact, region size is the 0-order shape moment; one can use higher-order shape moments for richer descriptions of shape.

$$\min_{\theta} \sum_{s} \sum_{i \in \Omega_s} \mathcal{L}(\mathbf{y}_s(i), \mathbf{p}(i, \theta)) + \gamma \mathcal{F}(\theta) \tag{2}$$

where γ is a positive constant and $\mathcal{F}$ a quadratic penalty, which takes the following form for the inequality constraints in (1):

$$\mathcal{F}(\theta) = \sum_{c=1}^{C} \sum_{t=1}^{T} [f_c(\mathbf{P}_t(\theta))]_+^2 \tag{3}$$

with $[x]_+ = \max(0, x)$ denoting the rectifier linear unit function.

3 Experiments

3.1 Experimental Set-Up

Dataset. The proposed method was evaluated on the publicly available MICCAI 2018 IVDM3Seg Challenge[3] dataset. This dataset contains 16 3D multimodal magnetic resonance (MR) scans of the lower spine, with their corresponding manual segmentations, collected from 8 subjects at two different stages in a study investigating intervertebral discs (IVD) degeneration. In our experiments, we employed the water (Wat) modality as the labeled source domain S and the in-phase (IP) modality as the unlabeled target domain T, and the setting is binary classification ($K = 2$). While 13 scans were used for training, the remaining 3 scans were employed for validation.

Constrained versus Adversarial Domain Adaptation. We compared our constrained DA model to the adversarial approach proposed in [20], which encourages the output space to be invariant across domains. To do so, the penalty $\mathcal{F}$ in (2) is replaced by an adversarial loss, which enforces the alignment between the distributions of source and target image segmentations. During training, pairs of images from the source and target domain are fed into the segmentation network. Then, a discriminator uses the generated masks as inputs and attempts to identify the domain from which the masks come from (source, or target). In this setting, we focused on a single-level adversarial learning for simplicity (see [20] for more details).

Diverse Levels of Supervision. We used the penalty term in (3) on the size of the target region (the IVDs) bounded by two prior values, which were estimated from the ground truth. This setting is later on referred to as *Constraint*. We also experimented with three different levels of tightness of the bounds, $\pm 10\%$, $\pm 50\%$ and $\pm 70\%$ of variations with respect to the actual size, so as to evaluate the behaviour of our method in the case of imprecise prior knowledge. In addition, we employed a model trained on the source as the lower baseline –without any adaptation strategy– and a model trained on the target data, referred to as *Oracle*, which serves as an upper bound.

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

Training and Implementation Details. As suggested in [20], we employ pairs of images from both domains, I_s and I_t, to train the deep models, which in our case correspond to the same 2D axial slice but from different modalities. For the segmentation network, we employ ENet [16], but any CNN segmentation network could be used. Regarding the DA adversarial approach, we employ the same segmentation network and include the discriminator proposed in [20]. Both the segmentation and the discrimination network were trained with Adam optimizer and a batch size of 1, for 100 epochs, and an initial learning rate of 5×10^{-4} and 10^{-4}, respectively. A baseline model trained on the source with full supervision was used as initialization. The γ parameter in (2) was set empirically to 2.5 in the proposed constrained adaptation model and to 0.1 in the adversarial approach.

Evaluation. In all our experiments, the Dice similarity coefficient (DSC) and the Hausdorff distance (HD) were employed as evaluation metrics to compare the different models.

3.2 Results

Quantitative metrics are reported in Table 1. First, we can observe that employing a model trained on source images to segment target images yields poor results, demonstrating the difficulty of CNNs to generalize well on a new domain. Adopting the adversarial strategy substantially improves the performance over the lower baseline, achieving a mean DSC of 65.3%. The proposed constrained DA models achieve a DSC value of 81.1%, 78.5% and 70.0% with tight ($Constraint_{10}$) and loose bounds ($Constraint_{50}$ and $Constraint_{70}$), respectively. This shows that, even with relaxed constraints, the proposed constrained DA model clearly outperforms the adversarial approach. Compared to the *Oracle*, the two best models –i.e., $Constraint_{10}$ and $Constraint_{50}$– reach 98% and 95% of its performance, demonstrating the efficiency of the proposed method and its robustness to the loosening of bounds. Regarding the HD values, we observe a similar pattern across the different models. Even though the adversarial approach reduces the HD to almost the half (1.67 pixels) compared to the lower baseline model (2.99 pixels), it is still far from the results obtained with our constrained models (1.10, 1.09 and 1.23 pixels). These findings are in line with the plots in Fig. 2, where the evolution of the training in terms of validation DSC is shown. In Fig. 2, *left* we can observe that the gap between the proposed and the adversarial approach holds during the whole training, with our constrained formulation yielding rapidly high validation Dice measures (first 20 epochs). This suggests that integrating the constraints help the learning process in domain adaptation.

Qualitative segmentations from the validation set are depicted in Fig. 3, from the easiest to the hardest subject. It can be observed that, if no adaptation is adopted, or even with the adversarial learning strategy, the network fails to successfully detect the 7 IVDs on all the subjects. While the adversarial approach segments 6 IVDs in the easiest subject (*top*), it is not able to correctly identify

Table 1. Quantitative comparisons of performance on the target domain for the different models.

	Source⟶Target					Target⟶Target
	No adaptation	Adversarial [20]	Constraint_{10}	Constraint_{50}	Constraint_{70}	Oracle
DSC	42.8 ± 5.29	65.3 ± 5.54	**81.1 ± 0.59**	78.5 ± 1.94	70.0 ± 4.11	82.9 ± 2.29
HD	2.99 ± 1.55	1.67 ± 1.64	1.10 ± 1.34	**1.09 ± 1.36**	1.23 ± 1.51	1.08 ± 1.35

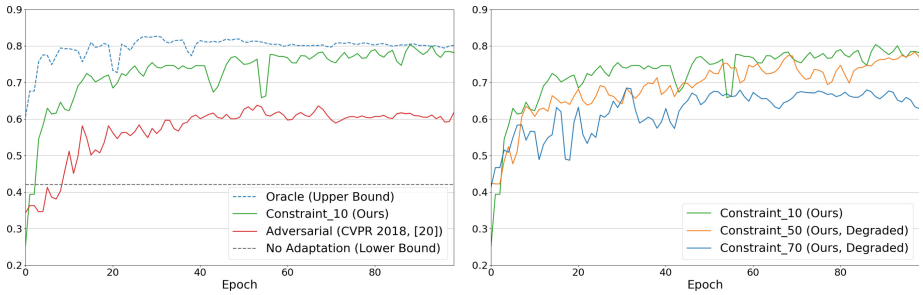

Fig. 2. Evolution of validation DSC over training for the different models. Comparison of the proposed model to the lower and upper bounds, as well as to the adversarial strategy is shown in the *left* figure, while an ablation study on the bounds is depicted in the *right*.

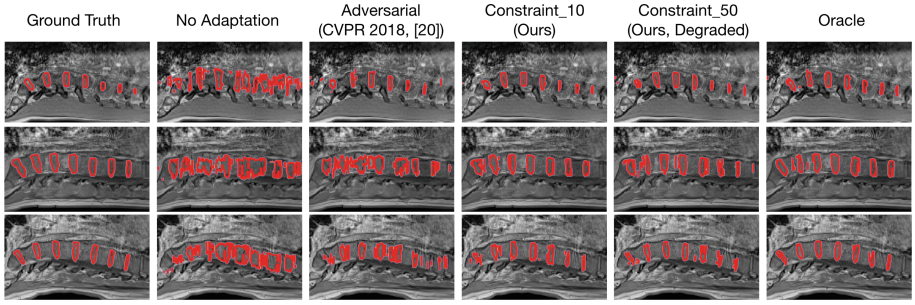

Fig. 3. Visual results in the validation set for several models. For better visibility results are depicted in the sagittal plane.

separate structures on harder cases. The segmentations achieved by the proposed constrained DA model present much better compactness and shape, where the 7 IVDs are distinguishable in all the subjects.

4 Conclusion

In this paper, we proposed a simple constrained formulation for domain adaptation in the context of semantic segmentation of medical images. Particularly, the proposed approach employs domain-invariant prior knowledge about the object

of interest, in the form of target size, which is derived from the source ground truth. Unlike adversarial strategies, which are based on two-step training, our method tackles the UDA problem with a single constrained loss, simplifying the adaptation of the segmentation network. As demonstrated in our experiments, the performance is significantly improved with respect to a state-of-the art adversarial method, and is comparable to the upper baseline supervised on the target. The proposed learning framework is very flexible, being applicable to any architecture and capable of incorporating a wide variety of constraints.

References

1. Berry, J.L., Moran, J.M., Berg, W.S., Steffee, A.D.: A morphometric study of human lumbar and selected thoracic vertebrae. Spine **12**(4), 362–367 (1987)
2. Bertsekas, D.P.: Nonlinear Programming. Athena Scientific, Belmont (1995)
3. Chen, Y., Li, W., Van Gool, L.: Road: reality oriented adaptation for semantic segmentation of urban scenes. In: CVPR (2018)
4. Cheplygina, V., de Bruijne, M., Pluim, J.P.W.: Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. MedIA **54**, 280–296 (2019)
5. Gholami, A., et al.: A novel domain adaptation framework for medical image segmentation. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 289–298. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_26
6. He, F.S., Liu, Y., Schwing, A.G., Peng, J.: Learning to play in a day: faster deep reinforcement learning by optimality tightening. In: ICLR (2017)
7. Hoffman, J., et al.: CYCADA: cycle-consistent adversarial domain adaptation. In: ICML (2018)
8. Hong, W., Wang, Z., Yang, M., Yuan, J.: Conditional generative adversarial network for structured domain adaptation. In: CVPR (2018)
9. Javanmardi, M., Tasdizen, T.: Domain adaptation for biomedical image segmentation using adversarial training. In: ISBI (2018)
10. Jia, Z., Huang, X., Chang, E.I., Xu, Y.: Constrained deep weak supervision for histopathology image segmentation. IEEE TMI **36**(11), 2376–2388 (2017)
11. Kamnitsas, K., et al.: Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 597–609. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_47
12. Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y., Ayed, I.B.: Constrained-CNN losses for weakly supervised segmentation. MedIA **54**, 88–99 (2019)
13. Klodt, M., Cremers, D.: segmentation with moment constraints. In: ICCV (2011)
14. Litjens, G., et al.: A survey on deep learning in medical image analysis. MedIA **42**, 60–88 (2017)
15. Márquez-Neila, P., et al.: Imposing hard constraints on deep networks: promises and limitations. In: CVPR Workshop on Negative Results (2017)
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., Krähenbühl, P., Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation. In: ICCV (2015)

18. Ren, J., Hacihaliloglu, I., Singer, E.A., Foran, D.J., Qi, X.: Adversarial domain adaptation for classification of prostate histopathology whole-slide images. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 201–209. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_23
19. Shu, R., Bui, H.H., Narui, H., Ermon, S.: A DIRTT-T approach to unsupervised domain adaptation. In: ICLR (2018)
20. Tsai, Y., Hung, W., Schulter, S., Sohn, K., Yang, M., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: CVPR (2018)
21. Tzeng, E., et al.: Adversarial discriminative domain adaptation. In: CVPR (2017)
22. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: ICCV (2017)
23. Zhang, Y., Miao, S., Mansi, T., Liao, R.: Task driven generative modeling for unsupervised domain adaptation: application to X-ray image segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 599–607. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_67
24. Zhao, H., et al.: Supervised segmentation of un-annotated retinal fundus images by synthesis. IEEE TMI **38**(1), 46–56 (2019)
25. 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)