Test-Time Adaptation with Shape Moments for Image Segmentation

Mathilde Bateson(✉), Herve Lombaert, and Ismail Ben Ayed

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

Abstract. Supervised learning is well-known to fail at generalization under distribution shifts. In typical clinical settings, the source data is inaccessible and the target distribution is represented with a handful of samples: adaptation can only happen at test time on a few (or even a single) subject(s). We investigate test-time *single-subject adaptation* for segmentation, and propose a *shape-guided entropy minimization* objective for tackling this task. During inference for a single testing subject, our loss is minimized with respect to the batch normalization's scale and bias parameters. We show the potential of integrating various shape priors to guide adaptation to plausible solutions, and validate our method in two challenging scenarios: MRI-to-CT adaptation of cardiac segmentation and cross-site adaptation of prostate segmentation. Our approach exhibits substantially better performances than the existing test-time adaptation methods. Even more surprisingly, it fares better than state-of-the-art domain adaptation methods, although it forgoes training on additional target data during adaptation. Our results question the usefulness of training on target data in segmentation adaptation, and points to the substantial effect of shape priors on test-time inference. Our framework can be readily used for integrating various priors and for adapting any segmentation network. The code is publicly available (https://github.com/mathilde-b/TTA).

Keywords: Test-time adaptation · Segmentation · Shape moments · Deep networks · Entropy minimization

1 Introduction

Deep neural networks have achieved state-of-the-art performances in various natural and medical-imaging problems [13]. However, they tend to under-perform when the test-image distribution is different from those seen during training. In medical imaging, this is due to, for instance, variations in imaging modalities and protocols, vendors, machines, clinical sites and subject populations. For semantic segmentation problems, labelling a large number of images for each different target distribution is impractical, time-consuming, and often impossible. To circumvent those impediments, methods learning robust representations with less supervision have triggered interest in medical imaging [5].

L. Wang et al. (Eds.): MICCAI 2022, LNCS 13434, pp. 736–745, 2022.
https://doi.org/10.1007/978-3-031-16440-8_70

This motivates *Domain Adaptation* (DA) methods: DA amounts to adapting a model trained on an annotated source domain to another target domain, with no or minimal new annotations for the latter. Popular strategies involve minimizing the discrepancy between source and target distributions in the feature or output spaces [18,19]; integrating a domain-specific module in the network [6]; translating images from one domain to the other [23]; or integrating a domain-discriminator module and penalizing its success in the loss function [19].

In medical applications, separating the source training and adaptation is critical for privacy and regulatory reasons, as the source and target data may come from different clinical sites. Therefore, it is crucial to develop adaptation methods, which neither assume access to the source data nor modify the pre-training stage. Standard DA methods, such as [6,18,19,23], do not comply with these restrictions. This has recently motivated *Source-Free Domain Adaptation* (SFDA) [3,9], a setting where the source data (neither the images nor the ground-truth masks) is unavailable during the training of the adaptation phase.

Evaluating SFDA methods consists in: (i) adapting on a dedicated training set *Tr* from the target domain; and (ii) measuring the generalization performance on an unseen test set *Te* in the target domain. However, emerging and recent *Test-Time Adaptation* (TTA) works, both in learning and vision [4,17,21] as well as in medical imaging [9,20], argue that this is not as useful as adapting directly to the test set *Te*. In various applications, access to the target distribution might not be possible. This is particularly common in medical image segmentation when only a single target-domain subject is available for test-time inference. In the context of image classification, the authors of [21] showed recently that simple adaptation of batch normalization's scale and bias parameters on a set of test-time samples can deal competitively with domain shifts.

With this context in mind, we propose a simple formulation for source-free and single-subject test-time adaptation of segmentation networks. During inference for a single testing subject, we optimize a loss integrating shape priors and the entropy of predictions with respect to the batch normalization's scale and bias parameters. Unlike the standard SFDA setting, we perform test-time adaptation on each subject separately, and forgo the use of target training set *Tr* during adaptation. Our setting is most similar to the image classification work in [21], which minimized a label-free entropy loss defined over test-time samples. Building on this entropy loss, we further guide segmentation adaptation with domain-invariant shape priors on the target regions, and show the substantial effect of such shape priors on TTA performances. We report comprehensive experiments and comparisons with state-of-the-art TTA, SFDA and DA methods, which show the effectiveness of our shape-guided entropy minimization in two different adaptation scenarios: cross-modality cardiac segmentation (from MRI to CT) and prostate segmentation in MRI across different sites. Our method exhibits substantially better performances than the existing TTA methods. Surprisingly, it also fares better than various state-of-the-art SFDA and DA methods, although it does not train on source and additional target data during adaptation, but just performs joint inference and adaptation on a single

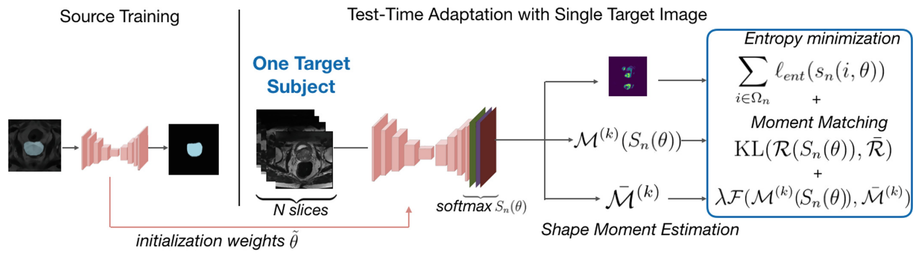
Source Training
Test-Time Adaptation with Single Target Image
One Target Subject
N slices
softmax $S_n(\theta)$
initialization weights $\tilde{\theta}$
$\mathcal{M}^{(k)}(S_n(\theta))$
$\bar{\mathcal{M}}^{(k)}$
Entropy minimization
$\sum_{i\in\Omega_n} \ell_{ent}(s_n(i,\theta))$
+
Moment Matching
$\mathrm{KL}(\mathcal{R}(S_n(\theta)), \bar{\mathcal{R}})$
+
$\lambda\mathcal{F}(\mathcal{M}^{(k)}(S_n(\theta)), \bar{\mathcal{M}}^{(k)})$
Shape Moment Estimation

Fig. 1. Overview of our framework for Test-Time Adaptation with Shape Moments: we leverage entropy minimization and shape priors to adapt a segmentation network on a single subject at test-time.

3D data point in the target domain. Our results and ablation studies question the usefulness of training on target set *Tr* during adaptation and points to the surprising and substantial effect of embedding shape priors during inference on domain-shifted testing data. Our framework can be readily used for integrating various priors and adapting any segmentation network at test times.

2 Method

We consider a set of M source images $I_m : \Omega_s \subset \mathbb{R}^2 \rightarrow \mathbb{R}$, $m = 1, \ldots, M$, and denote their ground-truth K-class segmentation for each pixel $i \in \Omega_s$ as a K-simplex vector $\mathbf{y}_m(i) = \left(y_m^{(1)}(i), \ldots, y_m^{(K)}(i)\right) \in \{0,1\}^K$. For each pixel i, its coordinates in the 2D space are represented by the tuple $\left(u_{(i)}, v_{(i)}\right) \in \mathbb{R}^2$.

Pre-training Phase. The network is first trained on the source domain only, by minimizing the cross-entropy loss with respect to network parameters θ:

$$\min_{\theta} \frac{1}{|\Omega_s|} \sum_{m=1}^{M} \ell\left(\mathbf{y}_m(i), \mathbf{s}_m(i,\theta)\right) \tag{1}$$

where $\mathbf{s}_m(i,\theta) = (s_m^{(1)}(i,\theta), \ldots, s_m^{(K)}(i,\theta)) \in [0,1]^K$ denotes the predicted softmax probability for class $k \in \{1, \ldots, K\}$.

Shape Moments and Descriptors. Shape moments are well-known in classical computer vision [15], and were recently shown useful in the different context of supervised training [10]. Each moment is parametrized by its orders $p, q \in \mathbb{N}$, and each order represents a different characteristic of the shape. Denote $I_n : \Omega_t \subset \mathbb{R}^2 \rightarrow \mathbb{R}$, $n = 1, \ldots, N$ the 2D slices of a subject in the target domain. For a given $p, q \in \mathbb{N}$ and class k, the shape moments of the segmentation prediction of an image I_n can be computed as follows from the softmax matrix $\mathrm{S}_n(\theta) = \left(\mathrm{s}_n^{(k)}(\theta)\right)_{k=1\ldots K}$:

$$\mu_{p,q}\left(\mathrm{s}_n^{(k)}(\theta)\right) = \sum_{i\in\Omega} s_n^{(k)}(i,\theta) u_{(i)}^p v_{(i)}^q$$

Central moments are derived from shape moments to guarantee translation invariance. They are computed as follows:

$$\bar{\mu}_{p,q}\left(\mathbf{s}_n^{(k)}(\theta)\right) = \sum_{i \in \Omega} s_n^k(i,\theta)\left(u_{(i)} - \bar{u}^{(k)}\right)^p \left(v_{(i)} - \bar{v}^{(k)}\right)^q .$$

where $\left(\frac{\mu_{1,0}(s_n^{(k)}(\theta))}{\mu_{0,0}(s_n^{(k)}(\theta))}, \frac{\mu_{0,1}(s_n^{(k)}(\theta))}{\mu_{0,0}(s_n^{(k)}(\theta))}\right)$ are the components of the centroid. We use the vectorized form onwards, e.g. $\mu_{p,q}\left(s_n(\theta)\right) = \left(\mu_{p,q}(s_n^{(1)}(\theta)), \ldots, \mu_{p,q}(s_n^{(K)}(\theta))\right)^\top$. Building from these definitions, we obtain 2D shape moments from the network predictions. We then derive the shape descriptors $\mathcal{R}, \mathcal{C}, \mathcal{D}$ defined in Table 1, which respectively inform on the size, position, and compactness of a shape.

Table 1. Examples of shape descriptors based on softmax predictions.

Shape Descriptor	Definition
Class-Ratio	$\mathcal{R}(s) := \frac{1}{\lvert\Omega_t\rvert}\mu_{0,0}\left(s\right)$
Centroid	$\mathcal{C}\left(s\right) := \left(\frac{\mu_{1,0}(s)}{\mu_{0,0}(s)}, \frac{\mu_{0,1}(s)}{\mu_{0,0}(s)}\right)$
Distance to Centroid	$\mathcal{D}\left(s\right) := \left(\sqrt[2]{\frac{\bar{\mu}_{2,0}(s)}{\mu_{0,0}(s)}}, \sqrt[2]{\frac{\bar{\mu}_{0,2}(s)}{\mu_{0,0}(s)}}\right)$

Test-Time Adaptation and Inference with Shape-Prior Constraints. Given a single new subject in the target domain composed of N 2D slices, $I_n : \Omega_t \subset \mathbb{R}^2 \rightarrow \mathbb{R}$, $n = 1, \ldots, N$, the first loss term in our adaptation phase is derived from [21], to encourage high confidence in the softmax predictions, by minimizing their weighted Shannon entropy: $\ell_{ent}(\mathbf{s}_n(i,\theta)) = -\sum_k \nu_k s_n^k(i,\theta) \log s_n^k(i,\theta)$, where $\nu_k, k = 1 \ldots K$, are class weights added to mitigate imbalanced class-ratios.

Ideally, to guide adaptation, for each slice I_n, we would penalize the deviations between the shape descriptors of the softmax predictions $S_n(\theta)$ and those corresponding to the ground truth $\mathbf{y_n}$. As the ground-truth labels are unavailable, instead, we estimate the shape descriptors using the predictions from the whole subject $\{S_n(\theta), n = 1, \ldots, N\}$, which we denote respectively $\bar{\mathcal{C}}, \bar{\mathcal{D}}$.

The first shape moment we leverage is the simplest: a zero-order class-ratio $\mathcal{R}$. Seeing these class ratios as distributions, we integrate a KL divergence with the Shannon entropy:

$$\mathcal{L}_{TTAS}(\theta) = \sum_n \frac{1}{|\Omega_n|} \sum_{i \in \Omega_t} \ell_{ent}(s_n(i,\theta)) + \mathrm{KL}(\mathcal{R}(S_n(\theta)), \bar{\mathcal{R}}). \tag{2}$$

It is worth noting that, unlike [2], which used a loss of the form in Eq. (2) for training on target data, here we use this term for inference on a test subject, as a part of our overall shape-based objective. Additionally, we integrate the centroid

($\mathcal{M} = \mathcal{C}$) and the distance to centroid ($\mathcal{M} = \mathcal{D}$) to further guide adaptation to plausible solutions:

$$\begin{aligned} \min_{\theta} \quad & \mathcal{L}_{TTAS}(\theta) \\ \text{s.t.} \quad & \left| \mathcal{M}^{(k)}(S_n(\theta)) - \bar{\mathcal{M}}^{(k)} \right| \leq 0.1, \quad k = \{2, \ldots, K\}, n = \{1, \ldots, N\}. \end{aligned} \tag{3}$$

Imposing such hard constraints is typically handled through the minimization of the Lagrangian dual in standard convex-optimization. As this is computationally intractable in deep networks, inequality constraints such as Eq. (3) are typically relaxed to soft penalties [7,8,11]. Therefore, we experiment with the integration of $\mathcal{C}$ and $\mathcal{D}$ through a quadratic penalty, leading to the following unconstrained objectives for joint test-time adaptation and inference:

$$\sum_n \frac{1}{|\Omega_t|} \sum_{i \in \Omega_n} \ell_{ent}(\mathbf{s}_n(i, \theta)) + \text{KL}(\mathcal{R}(S_n(\theta)), \bar{\mathcal{R}}) + \lambda \mathcal{F}(\mathcal{M}(S_n(\theta)), \bar{\mathcal{M}}), \tag{4}$$

where $\mathcal{F}$ is a quadratic penalty function corresponding to the relaxation of Eq. (3): $\mathcal{F}(m_1, m_2) = [m_1 - 0.9m_2]_+^2 + [1.1m_2 - m_1]_+^2$ and $[m]_+ = \max(0, m)$, with λ denoting a weighting hyper-parameter. Following recent TTA methods [9,21], we only optimize for the scale and bias parameters of batch normalization layers while the rest of the network is frozen. Figure 1 shows the overview of the proposed framework.

3 Experiments

3.1 Test-time Adaptation with Shape Descriptors

Heart Application. We employ the 2017 Multi-Modality Whole Heart Segmentation (MMWHS) Challenge dataset for cardiac segmentation [24]. The dataset consists of 20 MRI (source domain) and 20 CT volumes (target domain) of non-overlapping subjects, with their manual annotations of four cardiac structures: the Ascending Aorta (AA), the Left Atrium (LA), the Left Ventricle (LV) and the Myocardium (MYO). We employ the pre-processed data provided by [6]. The scans were normalized as zero mean and unit variance, and data augmentation based on affine transformations was performed. For the domain adaptation benchmark methods (DA and SFDA), we use the data split in [6]: 14 subjects for training, 2 for validation, and 4 for testing. Each subject has $N = 256$ slices.

Prostate Application. We employ the dataset from the publicly available NCI-ISBI 2013 Challenge[1]. It is composed of manually annotated T2-weighted MRI from two different sites: 30 samples from Boston Medical Center (source domain), and 30 samples from Radboud University Medical Center (target domain). For the DA and SFDA benchmark methods, 19 scans were used for training, one

[1] https://wiki.cancerimagingarchive.net.

for validation, and 10 scans for testing. We used the pre-processed dataset from [14], who resized each sample to 384 × 384 in axial plane, and normalized it to zero mean and unit variance. We employed data augmentation based on affine transformations on the source domain. Each subject has $N \in [15, 24]$ slices.

Benchmark Methods. Our first model denoted $TTAS_{\mathcal{RC}}$ constrains the class-ratio $\mathcal{R}$ and the centroid $\mathcal{C}$ using Eq. (4); similarly, $TTAS_{\mathcal{RD}}$ constrains $\mathcal{R}$ and the distance-to-centroid $\mathcal{D}$. We compare to two *TTA* methods: the method in [9], denoted *TTDAE*, where an auxiliary branch is used to denoise segmentation, and *Tent* [21], which is based on the following loss: $\min_\theta \sum_n \sum_{i\in\Omega_n} \ell_{ent}(\mathbf{s}_n(i,\theta))$. Note that *Tent* corresponds to performing an ablation of both shape moments terms in our loss. As an additional ablation study, $TTAS_{\mathcal{R}}$ is trained with the class-ratio matching loss in Eq. (2) only. We also compared to two *DA* methods based on class-ratio matching, *CDA* [1], and *CurDA* [22], and to the recent source-free domain adaptation (*SFDA*) method *AdaMI* in [2]. A model trained on the source only, *NoAdap*, was used as a lower bound. A model trained on the target domain with the cross-entropy loss, *Oracle*, served as an upper bound.

Estimating the Shape Descriptors. For the estimation of the class-ratio $\bar{\mathcal{R}}$, we employed the coarse estimation in [1], which is derived from anatomical knowledge available in the clinical literature. For $\mathcal{M} \in \{\mathcal{C}, \mathcal{D}\}$, we estimate the target shape descriptor from the network prediction masks $\hat{\mathbf{y}}_\mathbf{n}$ after each epoch: $\bar{\mathcal{M}}^{(k)} = \frac{1}{|V^k|}\sum_{v\in V^k} v$, with $V^k = \left\{\mathcal{M}^{(k)}(\hat{\mathbf{y}}_\mathbf{n}) \text{ if } \mathcal{R}^k(\hat{\mathbf{y}}_\mathbf{n}) > \epsilon^k, n = 1 \cdots N\right\}$.

Note that, for a fair comparison, we used exactly the same class-ratio priors and weak supervision employed in the benchmarks methods in [1,2,22]. Weak supervision takes the form of simple image-level tags by setting $\bar{\mathcal{R}}^{(k)} = \mathbf{0}$ and $\lambda = 0$ for the target images that do not contain structure k.

Training and Implementation Details. For all methods, the segmentation network employed was UNet [16]. A model trained on the source data with Eq. (1) for 150 epochs was used as initialization. Then, for TTA models, adaptation is performed on each test subject independently, without target training. Our model was initialized with Eq. (2) for 150 epochs, after which the additional shape constraint was added using Eq. (4) for 200 epochs. As there is no learning and validation set in the target domain, the hyper-parameters are set following those in the source training, and are fixed across experiments: we trained with the Adam optimizer [12], a batch size of $min(N, 22)$, an initial learning rate of 5×10^{-4}, a learning rate decay of 0.9 every 20 epochs, and a weight decay of 10^{-4}. The weights ν_k are calculated as: $\nu_k = \frac{\bar{\mathcal{R}}_k^{-1}}{\sum_k \bar{\mathcal{R}}_k^{-1}}$. We set $\lambda = 1 \times 10^{-4}$.

Evaluation. The 3D Dice similarity coefficient (DSC) and the 3D Average Surface distance (ASD) were used as evaluation metrics in our experiments.

3.2 Results and Discussion

Table 2 and Table 3 report quantitative metrics for the heart and prostate respectively. Among DA methods, the source-free *AdaMI* achieves the best DSC

Table 2. Test-time metrics on the cardiac dataset, for our method and various *Domain Adaptation* (DA), *Source Free Domain Adaptation* (SFDA) and *Test Time Adaptation* (TTA) methods.

Methods	DA	SFDA	TTA	DSC (%)					ASD (vox)				
				AA	LA	LV	Myo	Mean	AA	LA	LV	Myo	Mean
NoAdap (lower b.)				49.8	62.0	21.1	22.1	38.8	19.8	13.0	13.3	12.4	14.6
Oracle (upper b.)				91.9	88.3	91.0	85.8	89.2	3.1	3.4	3.6	2.2	3.0
CurDA [22]	✓	×	×	79.0	77.9	64.4	61.3	70.7	6.5	7.6	7.2	9.1	7.6
CDA [1]	✓	×	×	77.3	72.8	73.7	61.9	71.4	**4.1**	6.3	6.6	6.6	5.9
AdaMI [2]	×	✓	×	83.1	78.2	74.5	66.8	75.7	5.6	**4.2**	**5.7**	6.9	5.6
TTDAE [9]	×	×	✓	59.8	26.4	32.3	44.4	40.7	15.1	11.7	13.6	11.3	12.9
Tent [21]	×	×	✓	55.4	33.4	63.0	41.1	48.2	18.0	8.7	8.1	10.1	11.2
Proposed Method													
TTAS$_{\mathfrak{RC}}$ (Ours)	×	×	✓	**85.1**	**82.6**	**79.3**	**73.2**	**80.0**	5.6	4.3	6.1	**5.3**	**5.3**
TTAS$_{\mathfrak{RD}}$ (Ours)	×	×	✓	82.3	78.9	76.1	68.4	76.5	4.0	5.8	6.1	5.7	5.4
Ablation study													
TTAS$_{\mathfrak{R}}$	×	×	✓	78.9	77.7	74.8	65.3	74.2	5.2	4.9	7.0	7.6	6.2

improvement over the lower baseline *NoAdap*, with a mean DSC of 75.7% (cardiac) and 79.5% (prostate). Surprisingly though, in both applications, our method $TTAS_{\mathfrak{RD}}$ yields better scores: 76.5% DSC, 5.4 vox. ASD (cardiac) and 79.5% DSC, 3.9 vox. ASD (prostate); while $TTAS_{\mathfrak{RC}}$ achieves the best DSC across methods: 80.0% DSC and 5.3 vox. ASD (cardiac), 80.2% DSC and 3.79 ASD vox. (prostate). Finally, comparing to the TTA methods, both $TTAS_{\mathfrak{RC}}$ and $TTAS_{\mathfrak{RD}}$ widely outperform $TTADAE$, which yields 40.7% DSC, 12.9 vox. ASD (cardiac) and 73.2% DSC, 5.80 vox. ASD (prostate), and *Tent*, which

Table 3. Test-time metrics on the prostate dataset.

Methods	DA	SFDA	TTA	DSC (%)	ASD (vox)
NoAdap (lower bound)				67.2	10.60
Oracle (upper bound)				88.9	1.88
CurDA [22]	✓	×	×	76.3	3.93
CDA [1]	✓	×	×	77.9	**3.28**
AdaMI [2]	×	✓	×	79.5	3.92
TTDAE [9]	×	×	✓	73.2	5.80
Tent [21]	×	×	✓	68.7	5.87
Proposed Method					
TTAS$_{\mathfrak{RC}}$ (Ours)	×	×	✓	**80.2**	3.79
TTAS$_{\mathfrak{RD}}$ (Ours)	×	×	✓	79.5	3.90
Ablation study					
TTAS$_{\mathfrak{R}}$ (Ours)	×	×	✓	75.3	5.06

reaches 48.2% DSC, 11.2 vox. ASD (cardiac) and 68.7% DSC, 5.87 vox. ASD (prostate).

Qualitative segmentations are depicted in Fig. 2. These visuals results confirm that without adaptation, a model trained only on source data cannot properly segment the structures on the target images. The segmentation masks obtained using the TTA formulations $Tent$ [21], $TTADAE$ [9] only show little improvement. Both methods are unable to recover existing structures when the initialization $NoAdap$ fails to detect them (see fourth and fifth row, Fig. 2). On the contrary, those produced from our degraded model $TTAS_{\mathfrak{R}}$ show more regular edges and is closer to the ground truth. However, the improvement over $TTAS_{\mathfrak{R}}$ obtained by our two models $TTAS_{\mathfrak{RC}}$, $TTAS_{\mathfrak{RD}}$ is remarkable regarding the shape and position of each structures: the prediction masks show better centroid position (first row, Fig. 2, see LA and LV) and better compactness (third, fourth, fifth row, Fig. 2).

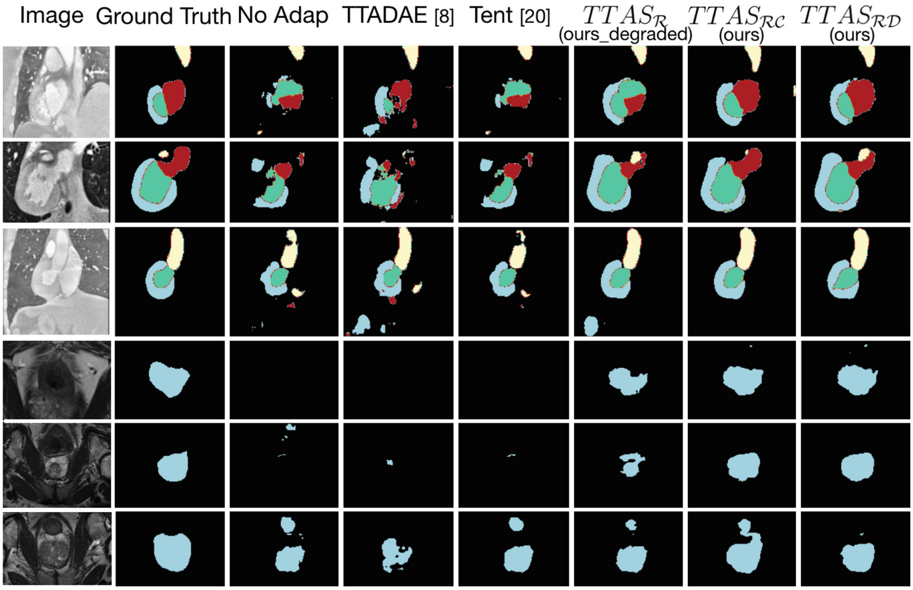

Fig. 2. Qualitative performance on cardiac images (top) and prostate images (bottom): examples of the segmentations achieved by our formulation ($TTAS_{\mathfrak{RC}}$, $TTAS_{\mathfrak{RD}}$), and benchmark TTA models. The cardiac structures of MYO, LA, LV and AA are depicted in blue, red, green and yellow respectively. (Color figure online)

4 Conclusion

In this paper, we proposed a simple formulation for *single-subject* test-time adaptation (TTA), which does not need access to the source data, nor the availability of a target training data. Our approach performs inference on a test subject by minimizing the entropy of predictions and a class-ratio prior over batchnorm parameters. To further guide adaptation, we integrate shape priors through penalty constraints. We validate our method on two challenging tasks, the MRI-to-CT adaptation of cardiac segmentation and the cross-site adaptation of prostate segmentation. Our formulation achieved better performances than state-of-the-art TTA methods, with a 31.8% (resp. 7.0%) DSC improvement on cardiac and prostate images respectively. Surprisingly, it also fares better than various state-of-the-art DA and SFDA methods. These results highlight the effectiveness of shape priors on test-time inference, and question the usefulness of training on target data in segmentation adaptation. Future work will involve the introduction of higher-order shape moments, as well as the integration of multiple shapes moments in the adaptation loss. Our test-time adaptation framework is straightforward to use with any segmentation network architecture.

Acknowledgments. This work is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Fonds de recherche du Quebec (FRQNT), the Canada Research Chair on Shape Analysis in Medical Imaging, and the ETS Research Chair on Artificial Intelligence in Medical Imaging.

References

1. Bateson, M., Dolz, J., Kervadec, H., Lombaert, H., Ben Ayed, I.: Constrained domain adaptation for image segmentation. IEEE Trans. Med. Imaging **40**(7), 326–334 (2021)
2. Bateson, M., Dolz, J., Kervadec, H., Lombaert, H., Ben Ayed, I.: Source-free domain adaptation for image segmentation (2021). https://arxiv.org/abs/2108.03152
3. Bateson, M., Kervadec, H., Dolz, J., Lombaert, H., Ben Ayed, I.: Source-relaxed domain adaptation for image segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 490–499. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_48
4. Boudiaf, M., Mueller, R., Ben Ayed, I., Bertinetto, L.: Parameter-free online test-time adaptation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
5. 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. Med. Image Anal. **54**, 280–296 (2019)
6. Dou, Q., et al.: Pnp-adanet: plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. IEEE Access **7**, 99065–99076 (2019)
7. He, F.S., Liu, Y., Schwing, A.G., Peng, J.: Learning to play in a day: faster deep reinforcement learning by optimality tightening. In: International Conference on Learning Representations (ICLR) (2017)
8. Jia, Z., Huang, X., Chang, E.I., Xu, Y.: Constrained deep weak supervision for histopathology image segmentation. IEEE Trans. Med. Imaging **36**(11), 2376–2388 (2017)
9. Karani, N., Erdil, E., Chaitanya, K., Konukoglu, E.: Test-time adaptable neural networks for robust medical image segmentation. Med. Image Anal. **68**, 101907 (2021)

10. Kervadec, H., Bahig, H., Létourneau-Guillon, L., Dolz, J., Ben Ayed, I.: Beyond pixel-wise supervision for segmentation: a few global shape descriptors might be surprisingly good! In: Medical Imaging with Deep Learning (MIDL) (2021)
11. 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)
12. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2014)
13. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
14. Liu, Q., Dou, Q., Heng, P.-A.: Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12262, pp. 475–485. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59713-9_46
15. Nosrati, M.S., Hamarneh, G.: Incorporating prior knowledge in medical image segmentation: a survey. arXiv preprint arXiv:1607.01092 (2016)
16. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
17. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: International Conference on Machine Learning (ICML) (2020)
18. Tsai, Y.H., et al.: Learning to adapt structured output space for semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
19. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
20. Varsavsky, T., Orbes-Arteaga, M., Sudre, C.H., Graham, M.S., Nachev, P., Cardoso, M.J.: Test-time unsupervised domain adaptation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 428–436. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_42
21. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (ICLR) (2021)
22. Zhang, Y., David, P., Foroosh, H., Gong, B.: A curriculum domain adaptation approach to the semantic segmentation of urban scenes. IEEE Trans. Pattern Anal. Mach. Intell. **42**, 1823–1841 (2020)
23. Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE/CVF Conference on Computer Vision (ICCV) (2017)
24. Zhuang, X., et al.: Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. Med. Image Anal. **58**, 101537 (2019)