OLVA: Optimal Latent Vector Alignment for Unsupervised Domain Adaptation in Medical Image Segmentation

Dawood Al Chanti, Diana Mateus

OLVA: *O*ptimal *L*atent *V*ector *A*lignment for Unsupervised Domain Adaptation in Medical Image Segmentation*

Dawood Al Chanti and Diana Mateus

École Centrale de Nantes, Laboratoire des Sciences du Numérique de Nantes LS2N, UMR CNRS 6004 Nantes, France.

Abstract. This paper addresses the domain shift problem for segmentation. As a solution, we propose OLVA, a novel and lightweight unsupervised domain adaptation method based on a Variational Auto-Encoder (VAE) and optimal transport (OT) theory. Thanks to the VAE, our model learns a shared cross-domain latent space that follows a normal distribution, which reduces the domain shift. To guarantee valid segmentations, our shared latent space is designed to model the shape rather than the intensity variations. We further rely on an OT loss to match and align the remaining discrepancy between the two domains in the latent space. We demonstrate OLVA's effectiveness for the segmentation of multiple cardiac structures on the public Multi-Modality Whole Heart Segmentation (MM-WHS) dataset, where the source domain consists of annotated 3D MR images and the unlabelled target domain of 3D CTs. Our results show remarkable improvements with an additional margin of 12.5% dice score over concurrent generative training approaches.

Keywords: Unsupervised domain adaptation · Cross modality · Variational Auto-Encoder · Optimal Transport · Cardiac Segmentation.

1 Introduction

Automatic segmentation from multi-modal images is essential for clinical assessment, diagnosis and treatment planning [1, 17]. Extensive literature has shown the effectiveness of convolutional neural networks in segmenting accurately cardiac structures [15, 18]. Yet, without proper adaptation these models fail when deployed across modalities, new subjects and different clinical sites, due to a domain shift [10] *e.g.* between the modalities' appearance as in Fig. 1. Designing models that can perform well across domains is key in medical applications where labels are scarce and expensive to obtain.

Semi-supervised and Unsupervised Domain Adaptation (UDA) approaches have been proposed to tackle the domain shift problem. The former assume a

* This work has been supported in part by the European Regional Development. Fund, the Pays de la Loire region on the Connect Talent scheme (MILCOM Project) and Nantes Métropole (Convention 2017-10470),

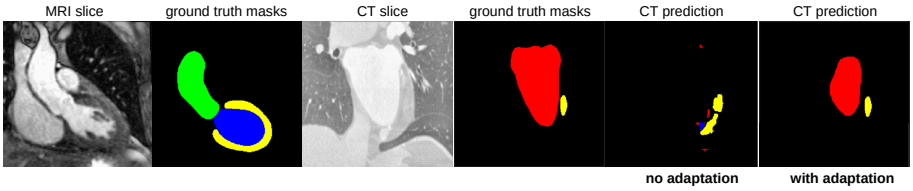

Fig. 1. The appearances of the cardiac structures look significantly different on MR and CT images. Both modalities share the same label space, yellow: left ventricle myocardium (LV-M), blue: left ventricle blood cavity (LV-B), red: left atrium blood cavity (LA-B), and green: ascending aorta (A-A). Bad prediction due to severe domain shift when no adaptation is considered. MM-WHS cardiac public database [25].

few labeled instances in the target domain can be used for joint-training with the source data [19]. The more ambitious UDA strategies [1, 3, 4, 17, 23, 24] assume no labels are available for the target domain. The core idea of UDA is to go through an adaption phase using a non-linear mapping to find a common domain-invariant representation or a latent space $\mathcal{Z}$. The domain shift in $\mathcal{Z}$ can be reduced by enforcing the two domains distributions to be closer via a certain loss (*e.g.* Maximum Mean Discrepancy [14]). Since $\mathcal{Z}$ is common to all domains who share the same label space, projected labeled source domain samples can be used to train a segmenter for all domains. In this paper, we deal with the problem of UDA for MR-CT cross-modality cardiac structure segmentation.

Related Work. Recent works on UDA for medical image segmentation rely on Generative Adversarial Networks [3, 4, 7, 17, 23, 24] to translate the appearance from one modality to the other using multiple discriminators and a pixel-wise cycle consistency loss. Despite their success, they: i) suffer from instabilities [2], ii) rely on complex architectures with more than 95 million parameters, iii) are prone to model collapse [16], and iv) may generate images outside the actual target domain [1]. To alleviate some of these limitations, Ouyang *et.al.* [17] combined adversarial networks with VAEs [13]. They exploited the VAEs constraint imposed on the latent space to match a prior distribution and experimentally validated that it reduces the domain shift when used as a shared space across domains. To encourage appearance-invariance, [17] deployed an adversarial loss guided by a cycle-consistency. The VAE model in [17] is complex as equipped with three encoders, three decoders, a segmenter, and a domain classier. Its loss function has six trade-off hyper-parameters to tune. Recently, Optimal Transport (OT) theory [22] jointly with deep learning methods was used by Ackaouy *et.al.* [1] where a joint cost measure combining both the distances at feature space of a deep 3D-Unet between the samples and a loss function measuring the discrepancy at the output space between the two domains is proposed. A limitation of Seg-JDOT is that it employs image patches to enable the generation of a higher number of samples and to avoid curse of dimensionality when optimizing for the transport plan γ.

Proposal. We present a novel and lightweight domain-invariant variational -segmentation auto-encoder model. We use the latent space of a VAE that is constrained to follow a prior normal distribution as a common space similar to [17] to reduce the domain shift. Then we exploit the geometry in $\mathcal{Z}$ for matching and aligning distances between probability distribution using OT theory by optimizing for a transport plan γ, similar to [1], to further shrink the remaining domain shift. Different from [17], who maximized the image likelihood, we directly learn a semantic latent representation that maximizes the label likelihood. Our idea is that the prior normal distribution has a limited capacity to handle intensity and shape variations, but it can be efficiently exploited for modelling shapes alone. This claim is supported by [18] who use a VAE as a post-processor on the top of a U-Net output to convert the erroneous U-Net predictions to anatomical plausible outputs. Conversely, we simultaneously perform anatomical plausible segmentation and partial alignment of the label-conditional distributions. Also, different from [1], i) we operate over the full image scale, ii) we bring the source and the target data closer to a normal distribution before solving for γ to guarantee its convergence and iii) we do not require the alignment of the label-conditional distributions at the label space.

***Our main contributions are*:** i) We reduce the domain shift between the source and the target domain by projecting them into a shared semantic latent space which is regularized to follow a prior normal distribution; ii) We address the remaining shift by aligning latent vectors from both domains using a discrepancy measure based on OT theory; iii) Different from a typical VAE which forces the latent space to model image intensity variations, we concentrate the limited capacity of the prior normal distribution to model the shape of the segmentation masks; iv) Our model is lightweight with 1.7 million parameters and easy to adapt for other clinical application; v) We validate our model on the MM-WHS public dataset and outperform state of the art methods by a margin of 12.5% dice score.

2 Method

Consider a labeled source domain dataset $\{X^s, Y^s\}_{s=1}^N$ with N images, and a target dataset $\{X^t\}_{t=1}^M$ with M images, but with unknown labels Y^t. The goal of UDA is to build a common space for X^s, X^t while using the source labels Y^s to guide a segmentation model to generalize across both domains. Here, we propose an Optimal Latent Vector Alignment (OLVA) method to learn a shared latent space that encodes all the structural information needed to generate image segmentation masks, regardless of the domain. We minimize the domain shift between the source and the target distributions in this latent space by pushing them close to a prior normal distribution with a VAE. An optimal transport discrepancy measure removes the remaining domain shift. Finally, a generative decoder guided by the source labels is trained to produce feasible semantic segmentation masks. A block diagram of the method is shown in Fig. 2.

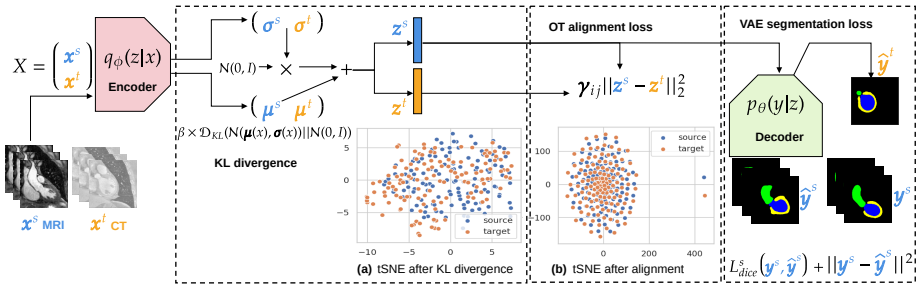

Fig. 2. OLVA: a generative encoder enforces the shared latent vectors of both source and target domain to follow a prior normal distribution (a), further aligned through an OT plan γ (b). A generative decoder is guided with the source domain labels to produce segmentation maps. We use t-distributed stochastic neighbor embedding (t-SNE) to map the latent vectors to a 2D space for visualization purposes only.

2.1 VAEs for segmentation

The goal of VAEs is to search for the best parameters ϕ^*, θ^* in order to sample a latent variable $\boldsymbol{z} \sim q_{\phi^*}(\boldsymbol{z}|\boldsymbol{x})$ whose distribution can be relatively simple such as isotropic Gaussian distribution and to generate a new sample $\hat{\boldsymbol{x}} \sim p_{\theta^*}(\boldsymbol{z}|\boldsymbol{x})$ as close as possible to the real observed data $\boldsymbol{x}$ such that $p^*_\theta(\boldsymbol{x}|\boldsymbol{z}) = p^*(x)$. The VAE loss formalized as in Eq. (1) enables an end-to-end training with a first term that maximize the marginal likelihood so that the generative model becomes better and a regularization term that minimize KL-divergence to better approximate $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ from the posterior $p_\theta(\boldsymbol{z})$. β is a trade-off parameter between the two terms. Commonly, a prior model $p(\boldsymbol{z})$ is set to normal distribution $\mathcal{N}(0; I)$ and the re-parametrization trick is applied to facilitate the sampling process as $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\mu_\phi(\boldsymbol{x}), \sigma_\phi(\boldsymbol{x})I)$. Thereby, $\mathcal{D}_{KL}$ become equivalent to $\frac{1}{2}\sum_{k=1}^{K}\left(1 + \log(\sigma_k^2) - \mu_k^2 - \sigma_k^2\right)$, K is the latent space dimension and $-\log p_\theta(\boldsymbol{x}|\boldsymbol{z})$ is conveniently replaced by a reconstruction loss $||\boldsymbol{x} - p_\theta(\boldsymbol{x}|\boldsymbol{z})||^2$.

$$\mathcal{L}_{vae}(\phi, \theta; \boldsymbol{x}) = \underset{z\sim q_\phi}{E}\left[-\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] + \beta\mathcal{D}_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z})), \tag{1}$$

To use VAEs for segmentation, we let $\boldsymbol{z}$ represent directly the latent space of the segmentation masks, since modelling shape alone is less complex than shape and intensity together. Moreover, VAEs lead to latent spaces that are continuous and structured (and not discrete as those of U-Net-like networks), facilitating interpolation. These choices will be determinant in producing valid segmentation masks that respect the anatomical variations of the source domain. In practice, our decoder acts as a predictive generative model for the conditional label distribution $p_\theta(\boldsymbol{y}|\boldsymbol{z})$ defined on the label space $\mathcal{Y}$. Using source domain data $\{X^s, Y^s\}$, our VAE segmentation loss is also guided by the soft dice loss $\mathcal{L}^s_{dice}$ to provide predictions of the segmentation maps as shown in Eq. (2.)

$$\mathcal{L}^s_{vae} = ||\boldsymbol{y}^s - p(\boldsymbol{y}^s|\boldsymbol{z}^s)||^2 + \beta\mathcal{D}^s_{KL}(q(\boldsymbol{z}^s|\boldsymbol{x}^s)||p_\theta(\boldsymbol{z}^s)) + \mathcal{L}^s_{dice}(\boldsymbol{y}^s, p(\boldsymbol{y}^s|\boldsymbol{z}^s)). \tag{2}$$

2.2 Optimal Transport for Latent Vector Alignment

To solve the domain adaptation problem within the segmentation task, we assume the existence of two distinct joint probability distributions $\mathcal{P}^s$ and $\mathcal{P}^t$ defined over a shared latent space $\mathcal{Z}$ and their marginal distributions (ζ^s, ζ^t) are defined over Ω (the set of all probability measures). Using $\mathcal{L}^s_{vae}(\phi, \theta; \boldsymbol{x}^s, \boldsymbol{y}^s)$, we regularize $\mathcal{P}^s$ to follow a normal distribution. To partially align $\mathcal{P}^s$ and $\mathcal{P}^t$, we argue that forcing both distribution to the same prior is beneficial. Therefore, we impose an additional cost function $\mathcal{D}^t_{KL}$ that measures the dissimilarity between latent vectors from the source $q(\boldsymbol{z}^s|\boldsymbol{x}^s)$ and the target $q(\boldsymbol{z}^t|\boldsymbol{x}^t)$ domains, with $\boldsymbol{z}^s, \boldsymbol{z}^t \in \mathcal{Z}$. Although necessary, this step is insufficient to completely align the two domains. The optimal transport (or Monge-Kantorovich) [1, 5, 9, 22] problem involves the matching of probability distributions defined over a geometric domain such as our latent semantic space. Here, using OT theory, we seek for a transportation plan matching distributions $\mathcal{P}^s$ and $\mathcal{P}^t$, which is equivalent to finding a probabilistic coupling, $\gamma_0 \in \prod(\mathcal{P}^s, \mathcal{P}^t)$, as shown in Eq. (3). To simultaneously align the latent space through a coupling γ_0 while optimizing for $q(\boldsymbol{z}|\boldsymbol{x})$, we adapt the Kantorovich OT formulation to the discrete case as in Eq. (4),

$$\gamma_0 = \underset{\prod(\mathcal{P}^s, \mathcal{P}^t)}{\arg\min} \int_{\Omega \times \Omega} \mathcal{D}(q(\boldsymbol{z^s}|\boldsymbol{x^s}); q(\boldsymbol{z^t}|\boldsymbol{x^t}))d\gamma(q(\boldsymbol{z^s}|\boldsymbol{x^s}); q(\boldsymbol{z^t}|\boldsymbol{x^t})). \quad (3)$$

$$\min_{\gamma \in \prod, q(\boldsymbol{z}|\boldsymbol{x})} \sum_{ij} \gamma_{ij} \mathcal{D}(q(\boldsymbol{z^s}|\boldsymbol{x^s}); q(\boldsymbol{z^t}|\boldsymbol{x^t})) + \beta \mathcal{D}^t_{KL}(q(\boldsymbol{z}^t|\boldsymbol{x}^t)||p_\theta(\boldsymbol{z^t})), \quad (4)$$

$$\min_{\gamma \in \prod, q(\boldsymbol{z}|\boldsymbol{x})} \sum_{ij} \gamma_{ij} \mathcal{D}(q(\boldsymbol{z^s}|\boldsymbol{x^s}); q(\boldsymbol{z}^t|\boldsymbol{x}^t)) + \beta \mathcal{D}^t_{KL}(q(\boldsymbol{z}^t|\boldsymbol{x}^t)||p_\theta(\boldsymbol{z}^t)) + \mathcal{L}^s_{vae} \quad (5)$$

where $\mathcal{D}(q(\boldsymbol{z^s}|\boldsymbol{x^s}); q(\boldsymbol{z^t}|\boldsymbol{x^t})) = \alpha||q(\boldsymbol{z^s}|\boldsymbol{x^s}) - q(\boldsymbol{z^t}|\boldsymbol{x^t})||_2^2$ is the squared Euclidean distance and $\beta \mathcal{D}^t_{KL}$ regularizes the target distribution. The final objective of OLVA is formulated in Eq. (5), which optimizes jointly for: i) an embedding function $q(\boldsymbol{z}|\boldsymbol{x})$ that maps both the source and the target domain to a semantic latent space $\mathcal{Z}$ regularized to follow normal distribution; ii) a transportation matrix γ that aligns similar semantic vectors $\boldsymbol{z}$ from both domains in the latent space; and iii) a predictive function $p(\boldsymbol{y}|\boldsymbol{z})$ for masks predictions.

2.3 Learning OLVA

With the formulation presented in in Eq. (5), our framework learns a common latent space that conveys aligned information for both the source and target domain. To solve Eq. (5), we use an alternating method [5]. Therefore, we optimize γ, with fixed $q(\boldsymbol{z}|\boldsymbol{x})$ and $p(\boldsymbol{y}|\boldsymbol{z})$, which reduces to the problem in Eq. (5) to solving a classic OT problem with cost matrix $C_{i,j} = \alpha||q(\boldsymbol{z}^s|\boldsymbol{x}^s) - q(\boldsymbol{z}^t|\boldsymbol{x}^t)||_2^2$. Then, we optimize $q(\boldsymbol{z}|\boldsymbol{x})$ and $p(\boldsymbol{y}|\boldsymbol{z})$, with fixed γ, this turns the problem in Eq. (5) to a standard deep learning problem. Similar to Damodoran *et.al.* [6], we solve the optimization problem with a stochastic approximation using mini-batches of size $m+n$ from the source and target domains respectively, which leads us to the

optimization problem presented in Eq. (6). The stochastic approximation yields a computationally feasible solution for both the OT and VAE. The discrepancy measure and the KL-Divergence regularization are computed at the latent space layer, while the segmentation loss uses the output layer.

$$\min_{q,p} \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}^{s}_{dice}(\boldsymbol{y}^s, p(\boldsymbol{y}^s|\boldsymbol{z}^s)) + \frac{1}{m}\sum_{i=1}^{m}\beta\mathcal{D}^{s}_{KL} + \frac{1}{n}\sum_{i=1}^{n}\beta\mathcal{D}^{t}_{KL} + \frac{1}{m}\sum_{i=1}^{m}||\boldsymbol{y}^s - p(\boldsymbol{y}^s|\boldsymbol{z}^s)||^2 + \min_{\gamma\in\Gamma(\zeta^s,\zeta^t)}\sum_{i,j}^{m+n}\gamma_{i,j}\alpha||q(\boldsymbol{z}^s|\boldsymbol{x}^s) - q(\boldsymbol{z}^t|\boldsymbol{x}^t)||^2\right] \tag{6}$$

Architecture and implementation details: OLVA accept batches containing 128 source and 128 target samples. The input dimension is $256 \times 256 \times 3$. The encoder is composed of five convolutional layers, with stride by 2 for downsampling, and with a leaky rectified linear unit (lrelu) activation, with a leakage rate of 0.3. The number of feature maps is successively $32, 32, 64, 64$, and 64. The last convolutional is flattened and mapped using a linear fully connected layer into two vectors (μ, σ), each composed of $K = 128$ features followed by a dropout of rate 0.3. A latent vector $\boldsymbol{z}$ is generate as $\mu + \sigma \odot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ and given as an input to the decoder. The decoder is composed of five up-convolutional layers, with a lrelu activation, each composed of $64, 64, 32, 32$, and 4 feature maps. The output layer with a sigmoid activation provides a mask of shape $256 \times 256 \times 4$. A learning rate of 0.0001 is used with Adam optimizer. Using the validation set we experimentally tuned, $\alpha = 10$ to focus more on the alignment loss and $\beta = 0.1$. The total number of iterations is 10,000.

3 Experiments and Results

We use the public MM-WHS dataset [25] for cardiac segmentation consisting of 20 MR ($\sim$ 128 slices) and 20 CT ($\sim$ 256 slices) unpaired and multi-site images from 40 patients. We followed the state-of-the-art data processing, domain adaptation protocol and evaluation metrics [11, 17, 7, 3, 4]. For data processing, we use the coronal view slices, cropped to 256×256 and normalized to zero mean and unit variance. To consider contextual information three adjacent slices ($256 \times 256 \times 3$) were stacked at the input and the middle slice label was used as the ground truth. Data augmentation included rotation, scaling, and affine transformations. A total of 11998 MR and 9598 sub-volumes were generated (each $256 \times 256 \times 3$). For domain adaptation, we randomly split each modality into training (16 subject) and testing (4 subjects). We use MR as a source domain, with 9599 sub-volumes for training and 2399 for validation. We set CT as a the target domain, with 8399 sub-volumes for training and 1199 for evaluation. We report the performance in terms of Dice Similarity Coefficient (DSC) and the Average Symmetric Surface Distance (ASSD).

Experimental Settings and Results: We consider four experimental settings. *First*, we compare our model without optimal transport and trained with

full supervision over the CT images (oracle VAE). We compare this setting against a U-Net [20] to show how our VAE constrains the shape of the predictions to be valid. The oracle VAE also serves as an upper-bound baseline. *Second*, to illustrate the domain shift problem, we consider the situation when no adaptation is performed, thereby, we train VAE-0 and U-Net-0 over MR images and evaluated them over CT images. *Third*, we consider the SOA setting, in which 16 labeled and 16 unlabeled source and target sequences are used (OLVA-16). We compare this setting with four SOA methods for medical UDA: PnP-AdaNet [7], SIFA [3], Synseg-net [11] and Seg-DJOT [1]. We also compare to two SOA methods for natural image UDA: CycleGAN [12] and AdaOutput [21]. *Fourth*, we consider a more ambitious scenario where we assume that only one unlabeled target sequence is available (OLVA-1). Therefore, we randomly draw one scan from the target set. To train OLVA-1 and to avoid overfitting, we fix all the model parameters and only the fully connected layer is retrained using loss evaluated at the latent space of Eq. 6. We also perform an experiment where an auxiliary reconstruction task [8] is integrated (OLVA-R-1). The quantitative and qualitative results are presented in Table 1 and Fig. 3

Discussion: Table 1 shows that our supervised baseline method outperforms the U-Net, achieving a high DSC and more importantly producing valid cardiac shape predictions as seen in Fig. 3, and as reflected by the ASSD score. When no adaptation is considered, VAE-0 achieves 49% DSC, while U-Net achieved only 15%. As our VAE-0 pushes the latent semantic features to be close to normal distribution, it partially aligns the marginal distributions. In the UDA setting, OLVA-16 outperforms the SOA's best results by an additional 12.5% in DSC! and having minimal erroneous prediction as seen in Fig. 3, with average ASSD of 0.31 mm. Considering the target data scarcity UDA Setting, OLVA-1 achieved the second best results after DECM-1 with an 8% DSC difference. The results of OLVA-1 when the reconstruction auxiliary task is introduced (OLVA-R-1) reduce the gap to 5% at the price of increasing the model complexity. The second-place is honorable, comparing the 1.7 million parameters of OLVA-1 with the more than 95 million parameters of DECM-1, and considering the quality of predictions as reflected by the ASSD. We also examine OLVA's performance when trained with randomly sampled 5 and 8 targets sequences. OLVA-5 achieves similar

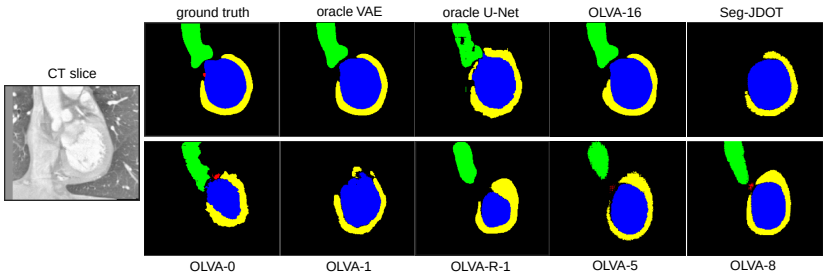

Fig. 3. Qualitative results of adaptation performances on segmentation.

Table 1. Performances of UDA from MR to CT images under different settings. Postfix -0, -1 or -16 after names of each method indicate the number of unlabelled target scans used for training. We mark in bold the best results and we underline the second best.

	DSC Score					**ASSD Score (mm)**				
Methods	LV-M	LA-B	LV-B	A-A	**avg**	LV-M	LA-B	LV-B	A-A	**avg**
oracle U-Net	0.83	0.89	0.92	0.93	0.89	0.38	0.39	0.28	0.31	0.34
oracle VAE	**0.95**	**0.97**	**0.97**	**0.96**	**0.96**	**0.06**	**0.04**	**0.03**	**0.05**	**0.05**
U-Net-0	0.10	0.27	0.02	0.24	0.15	36.0	19.4	48.6	31.9	26.2
VAE-0	**0.41**	**0.51**	**0.60**	**0.48**	**0.49**	**2.51**	**2.21**	**2.44**	**2.95**	**2.53**
OLVA-16	**0.79**	**0.87**	**0.88**	**0.88**	**0.85**	**0.56**	**0.53**	**0.37**	**0.45**	**0.31**
Seg-DJOT-16	0.57	0.60	0.57	0.62	0.59	3.64	3.62	3.85	5.20	4.07
SIFA-16	0.58	0.76	0.76	0.81	0.73	3.44	3.83	3.30	2.64	3.32
Pnp-AdaNet-16	0.50	0.77	0.60	0.79	0.66	10.2	4.04	8.60	2.28	6.22
SynSeg-Net-16	0.41	0.69	0.52	0.72	0.58	4.60	3.80	3.40	5.60	4.35
AdaOutput-16	0.43	0.76	0.54	0.65	0.59	4.68	2.89	3.10	6.15	4.20
CycleGAN-16	0.28	0.75	0.52	0.73	0.57	4.85	6.20	3.92	5.54	5.30
OLVA-1	0.58	0.69	0.64	0.67	0.64	2.10	1.95	1.85	**2.30**	2.05
OLVA-R-1	**0.68**	0.70	0.78	0.60	0.69	**1.89**	**1.88**	**1.51**	2.43	**1.92**
DECM-1	0.60	**0.78**	0.71	**0.78**	**0.72**	7.37	3.87	6.44	2.77	5.11
Seg-DJOT-1	0.19	0.25	0.21	0.20	0.21	9.64	13.7	8.18	10.3	10.4
SIFA-1	0.39	0.53	**0.80**	0.62	0.62	12.8	4.12	7.70	2.72	6.84
Pnp-AdaNet-1	0.29	0.48	0.33	0.58	0.25	25.1	27.1	27.7	7.14	21.8

performance to DCEM-1, while OLVA-8 achieves better DSC score 79%. As for further ablation studies, we change the latent dimension to 64, 256 and 512. With $K = 64$, a degradation in the source domain performance was observed, yielding an average DSC score of 79%. With $K = 256$, similar performances to $K = 128$ is achieved. When $K = 512$, a degradation in the performance over the source and the target domain is observed, leading to 69.6% target DSC score for OLVA-16. This degradation is expected as optimizing for γ requires a reasonable number of samples which grows with K's dimensionality [1].

4 Conclusion

To improve the applicability of deep learning model on new modality where it is expensive to acquire expert annotations, unsupervised domain adaptation represents a central solution. In this paper, we tackle the problem of unsupervised cross-modality medical image segmentation with a novel framework that jointly integrates VAE and OT theory to solve UDA problem. OLVA is a simple, efficient and lightweight model, which makes it practical to deploy in real-life without requiring a machine with huge computational resources. The usability of our method can be integrated within other learning regimes, for instance, a weakly-supervised model where sparse annotation of biomedical volumetric data are available and the aim would be to leverage the rest of the unlabeled data by matching them with the available labeled set. Future work will address the

problem of building a general segmenter where the adaptation from one task to another is done with minimal task-specific information and to leverage other tasks information.

References

1. Ackaouy, A., Courty, N., Vallée, E., Commowick, O., Barillot, C., Galassi, F.: Unsupervised domain adaptation with optimal transport in multi-site segmentation of multiple sclerosis lesions from mri data. Frontiers in computational neuroscience **14**, 19 (2020)
2. Arjovsky, M., Bottou, L.: Towards principled methods for training generative adversarial networks. In International Conference on Learning Representations (2017)
3. Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 865–872 (2019)
4. Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. IEEE transactions on medical imaging **39**, 2494–2505 (2020)
5. Courty, N., Flamary, R., Tuia, D., Rakotomamonjy, A.: Optimal transport for domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**, 1853–1865 (2017). https://doi.org/10.1109/TPAMI.2016.2615921
6. Damodaran, B.B., Kellenberger, B., Flamary, R., Tuia, D., Courty, N.: Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 447–463 (2018)
7. Dou, Q., Ouyang, C., Chen, C., Chen, H., Glocker, B., Zhuang, X., Heng, P.A.: Pnp-adanet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation. arXiv preprint arXiv:1812.07907 (2018)
8. Duque, V.G., Al Chanti, D., Crouzier, M., Nordez, A., Lacourpaille, L., Mateus, D.: Spatio-temporal consistency and negative label transfer for 3d freehand us segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 710–720. Springer (2020)
9. Gopalan, R., Li, R., Chellappa, R.: Domain adaptation for object recognition: An unsupervised approach. In: 2011 international conference on computer vision. pp. 999–1006. IEEE (2011)
10. Heimann, T., Mountney, P., John, M., Ionasec, R.: Learning without labeling: Domain adaptation for ultrasound transducer localization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 49–56. Springer (2013)
11. Huo, Y., Xu, Z., Moon, H., Bao, S., Assad, A., Moyo, T.K., Savona, M.R., Abramson, R.G., Landman, B.A.: Synseg-net: Synthetic segmentation without target modality ground truth. IEEE transactions on medical imaging **38**(4), 1016–1025 (2018)
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
13. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2014)

14. Kumagai, A., Iwata, T.: Unsupervised domain adaptation by matching distributions based on the maximum mean discrepancy via unilateral transformations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 4106–4113 (2019)
15. Li, F., Li, W., Qin, S., Wang, L.: Mdfa-net: Multiscale dual-path feature aggregation network for cardiac segmentation on multi-sequence cardiac mr. Knowledge-Based Systems p. 106776 (2021)
16. Liu, K., Tang, W., Zhou, F., Qiu, G.: Spectral regularization for combating mode collapse in gans. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6382–6390 (2019)
17. Ouyang, C., Kamnitsas, K., Biffi, C., Duan, J., Rueckert, D.: Data efficient unsupervised domain adaptation for cross-modality image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 669–677. Springer (2019)
18. Painchaud, N., Skandarani, Y., Judge, T., Bernard, O., Lalande, A., Jodoin, P.M.: Cardiac segmentation with strong anatomical guarantees. IEEE Transactions on Medical Imaging **39**(11), 3703–3713 (2020)
19. Puybareau, É., Zhao, Z., Khoudli, Y., Carlinet, E., Xu, Y., Lacotte, J., Géraud, T.: Left atrial segmentation in a few seconds using fully convolutional network and transfer learning. In: International Workshop on Statistical Atlases and Computational Models of the Heart. pp. 339–347. Springer (2018)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
21. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7472–7481 (2018)
22. Villani, C.: Optimal transport: old and new, vol. 338. Springer Science & Business Media (2008)
23. Wu, F., Zhuang, X.: Cf distance: A new domain discrepancy metric and application to explicit domain adaptation for cross-modality cardiac image segmentation. IEEE Transactions on Medical Imaging **39**, 4274–4285 (2020)
24. Yang, J., Dvornek, N.C., Zhang, F., Zhuang, J., Chapiro, J., Lin, M., Duncan, J.S.: Domain-agnostic learning with anatomy-consistent embedding for cross-modality liver segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
25. Zhuang, X., Shen, J.: Multi-scale patch and multi-modality atlases for whole heart segmentation of mri. Medical image analysis **31**, 77–87 (2016)