Supplementary Material Document: Unsupervised Domain Adaptation with Contrastive Margin-Enhanced VAE for Cross-Modality Cardiac Segmentation

Lihong Qiao, Rui Wang, Bin Xiao, Shuyu Cheng, Weisheng Li, *Member, IEEE,* Xinbo Gao, *Senior, IEEE*
and Baiying Lei, *Senior, IEEE*

I. Proof of the Lower Bound

In this section, we give the proof of Eq.(4) in the body of the paper. For common VAE, its ultimate goal is to find a probability distribution $p\left(\theta\right)$ parameterized by θ and maximize the probability of occurrence of all pairs of (x, y) in the data combined. However, the distribution p_θ of the real data is nebulous and hard to perceive, so we required a quantifiable parametric model q_ϕ to approximate the VAE (using neural networks easily [1]) to fit such a hard-to-measure probability distribution. Thus, as with most variational self-encoders, we introduce a new distribution q_ϕ to approximate the elusive p_θ.

Assumption: For a immaculate encoder, the mapping from image to manifold should be deterministic and related only to the data and its distribution, so as to ensure that the manifold is obtained by the encoder mapping is representative. So we assume that y and z be conditionally independent on x for distribution q_ϕ.

First, rewriting $\log p_\theta\left(x, y\right)$ [1] in VAE as follows:

$$\begin{aligned}
&\log p_\theta(x, y)\\
&= \int q_\phi(z \mid x, y) \log\left[\frac{q_\phi(z|x,y)}{p_\theta(z|x,y)} \cdot \frac{p_\theta(z)}{q_\phi(z|x,y)} \cdot p_\theta(x, y \mid z)\right] dz\\
&= \int q_\phi(z \mid x, y) \log\left[\frac{q_\phi(z|x,y)}{\frac{p_\theta(z,y|x)}{p_\theta(y|x)}} \cdot \frac{p_\theta(z)}{q_\phi(z|x,y)} \cdot p_\theta(x, y \mid z)\right] dz\\
&= \int q_\phi(z \mid x, y) \log\left[\frac{q_\phi(z|x,y)}{p_\theta(z,y|x)} \cdot \frac{p_\theta(z)}{q_\phi(z|x,y)} \cdot p_\theta(y \mid x) \cdot p_\theta(x, y \mid z)\right] dz\\
&= D_{KL}\left(q_\phi(z, y \mid x) \mid p_\theta(z \mid x, y)\right) - D_{KL}\left(q_\phi(z \mid x, y) \mid p_\theta(z)\right) +\\
&\quad E_{q_\phi(z|x,y)} \log\left[p_\theta(y \mid x)\right] + E_{q_\phi(z|x,y)} \log\left[p_\theta(x, y \mid z)\right].
\end{aligned}$$

since $D_{KL}\left(q_\phi\left(z, y \mid x\right) | p_\theta\left(z \mid x, y\right)\right)$ is nonnegative, we can rewrite the above equation as:

$$\begin{aligned}
&\log p_\theta(x, y)\\
&\geq -D_{KL}\left(q_\phi(z \mid x, y) \mid p_\theta(z)\right) +\\
&\quad E_{q_\phi(z|x)} \log\left[p_\theta(y \mid x)\right] + E_{q_\phi(z|x)} \log\left[p_\theta(x, y \mid z)\right] \qquad (1)\\
&= -D_{KL}\left(q_\phi(z \mid x, y) \mid p_\theta(z)\right) + E_{q_\phi(z|x)} \log\left[p_\theta(y \mid x)\right] +\\
&\quad E_{q_\phi(z|x)} \log\left[p_\theta(y \mid z)\right] + E_{q_\phi(z|x)} \log\left[p_\theta(x \mid y, z)\right]
\end{aligned}$$

The equality holds by $p_\theta\left(x, y \mid z\right) = p_\theta\left(y \mid z\right) \cdot p_\theta\left(x \mid y, z\right)$. Fig. 1 illustrates the probability graph of the proposed VAE.

II. Network

As Fig. 1 illustrates, ME-VAE is comprised of three modules, the encoder for approximation $q_\phi\left(z \mid x\right)$, the margin enhanced module, which can also be set for different task goals, integrated into the decoder modeling the adaptive task term $E_{q_\phi(z|x)} \log\left[p_\theta\left(y \mid x\right)\right]$. While the decoder modeling the reconstruction term $E_{q_\phi(z|x)} \log\left[p_\theta\left(x \mid y, z\right)\right]$, and the segmentor modeling from the prediction term $E_{q_\phi(z|x)} \log\left[p_\theta\left(y \mid z\right)\right]$.

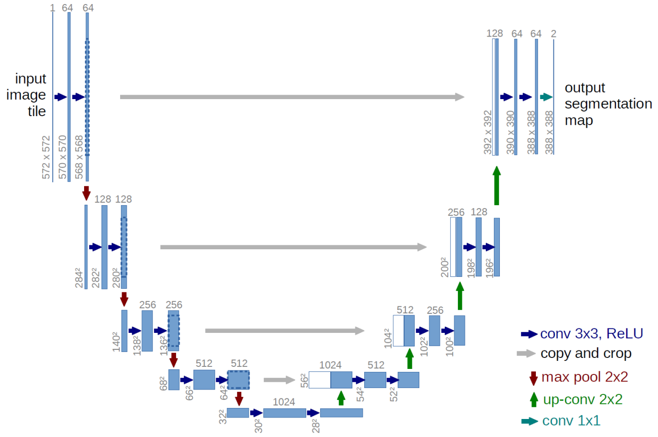

Fig. 2: The primitive U-net [2].

Our model is based on a VAE where the encoder of the VAE is a U-shape network, while the decoder consists

This work was supported in part by the National Key Research and Development Project under Grant 2019YFE0110800, in part by the National Natural Science Foundation of China under Grant 61806032 and Grant 61976031, in part by the National Major Scientific Research Instrument Development Project of China under Grant 62027827, in part by the Scientific and Technological Key Research Program of Chongqing Municipal Education Commission under Grant KJQN20220. (Corresponding author: Bin Xiao.)

Lihong Qiao, Rui Wang, Bin Xiao, Shuyu Cheng, Weisheng Li and Xinbo Gao are with the Department of Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail:{qiaolh, s210201096@stu(Rui Wang), xiaobin, shuyc, liws, gaoxb}@cqupt.edu.cn.).

Baiying Lei are with School of Biomedical Engineering, Shenzhen University, National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, Shenzhen, China (e-mail: leiby@szu.edu.cn).

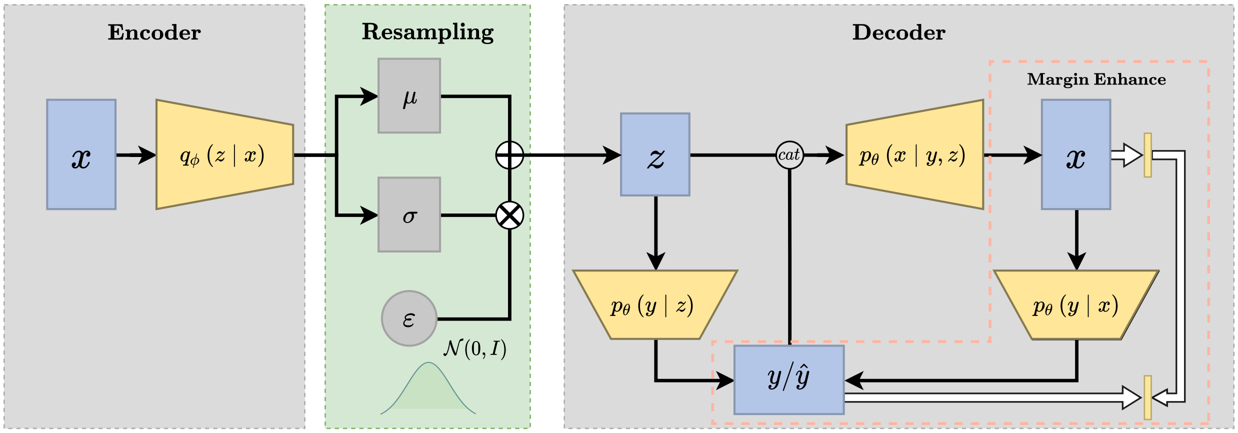

Fig. 1: **The illustration of our ME-VAE's probability graph architecture.**

of a segmentator and a conventional reconstructor. In our implementation, we use a five-layer U-net as the encoder in pursuit of multi-scale heart graphics information. Fig. 2 shows the primitive U-net, the basic structure of our encoder is identical to it except that the resolution of the input image changes with the dataset.

For the decoders, we set up separate decoders for the reconstruction of the different resolution images. Each decoder reconstructs the image using the features in the upsampling layer of the encoder at different resolutions combined with the segmentation results given by the segmentation head as input. The final combination of reconstruction effect metrics from multiple scales on the reconstruction term $E_{q_\phi(z|x)} \log \left[p_\theta(x \mid y, z)\right]$ approximation ability.

For the segmentation head, which is implemented through a convolutional layer that changes the dimensionality of the feature into $R^{B \times N_c \times \frac{W}{2^n} \times \frac{H}{2^n}}$ and eventually into a one-hot thermal form, we also set up a multi-scale segmentation head, which is implemented by cropping the source domain image and their labels in the same position and finally up-sampling to a uniform scale weighting as the final segmentation result. Warningly, the loss is calculated separately for each scale of the segmentation head during training.

III. EXPERIMENT SUPPLEMENT AND DETAILS

CMEVAE is being compared with several state-of-the-art unsupervised domain adaptation methods to demonstrate the superiority of CMEVAE in the MSCMR-Seg. The following unsupervised domain adaptation methods are included for comparison.

- SIFA [3] is a well-known unsupervised bi-directional cardiac domain adaption method that conducts co-alignment from the perspective of the image domain and feature domain by a shared encoder.
- CFDnet [4] performs a new metric based on characteristic functions of distributions that enables explicit domain adaptation rather than implicit domain adaptation via adversarial learning.
- PMKNet [5] distills the segmentation capability of unimodal images of student models via a teacher model trained with existing multimodal data and incorporates an image-level encoding of structural information to ensure the validity of the knowledge.
- UDA-VAE++ [6] develops a structure mutual information estimation block to maximize the mutual information between the reconstruction and segmentation tasks.
- MTL [7] is a multi-task model that proposes two novel self-supervised tasks of encoding spatial and morphological appearance information of cardiac images in a multi-task learning framework.
- DDFseg [8] disentangles domains into the domain-invariant features and orthogonal domain-specific features. Next, a GAN with multiple discriminators uses both types of information to segment across modalities.

A. Discussion of Convergence

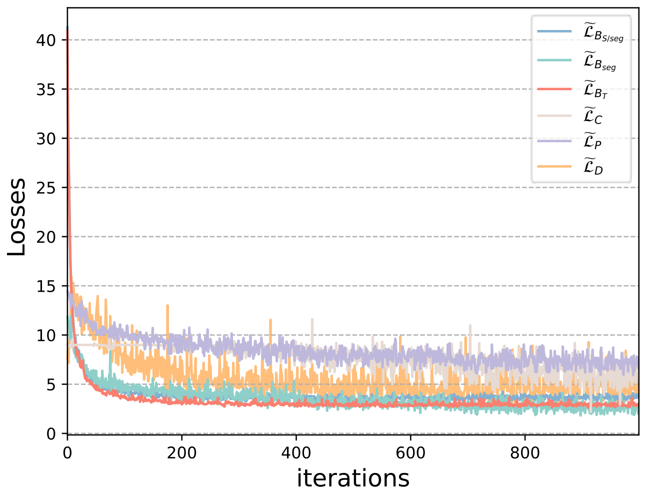

Fig. 3: The losses of CMEVAE during training on MS-CMR.

The loss term of CMEVAE during unidirectional training is further plotted in Fig. 3. To boost the significance of the contribution of individual terms, the segmentation loss that contributes significantly to the source domain of the model optimization is divided out and noted as $\widetilde{\mathcal{L}}_{B_{seg}}$. As shown in Fig. 3, the variational inference terms, $\widetilde{\mathcal{L}}_{B_{S/seg}}$ and $\widetilde{\mathcal{L}}_{B_T}$, decrease rapidly and converge quickly. In addition, during the initial phase of training, certain disorientation occurs in $\widetilde{\mathcal{L}}_C$ because the distinction between modalities is unaware, and with the help of source domain labels, $\widetilde{\mathcal{L}}_{B_{seg}}$, $\widetilde{\mathcal{L}}_{P}$ and $\widetilde{\mathcal{L}}_{C}$ continuously help the CMEVAE to learn comprehensive and detailed visual features.

B. Experimental Setup

MS-CMRSeg dataset[1]: The bSSFP-CMR cover the entire ventricle from the apex to the base of the mitral valve. The LGE-CMR cover the main body of the ventricles. The main processing involved shuffling BSSFP CMR and LGE CMR images collected from the same subject so that they were unpaired. For experiments, all images were intensity normalized, resized to be the same resolution of 1.0×1.0 mm and cropped with an ROI of 192×192 pixel. Details are presented in the baseline [9].

Myops dataset[2]: For preprocessing, due to the imbalance of foreground and background, we count the coordinate range of the targets in the training data. We expanded 30 voxels along each dimension, and crop the most valuable 320×320 area of the images accordingly, for details of the process refer to [3] [10]. In experiments, all images were intensity normalized and cropped with an ROI of 192×192 pixel to ensure experimental coherence. Among all 25 training cases, we randomly selected 20 as the training set and the rest were used to validate the training.

CT-MR dataset[4]: Unlike the project, we did not use pseudo-labels as an additional adjunct, but only used the labeled 20 pairs of cases for training. In experiments, all images were intensity normalized and cropped with an ROI of 192×192 pixel to ensure experimental coherence.

For the memory bank in CPMC, we stored 50 samples and a prototype for each point class at three scales and used all the stored samples in each contrast round, and selected the point with the lowest similarity to the prototype to update the memory bank. All models were trained using one single NVIDIA 3090 24GB GPU for 30 epochs in direction 1 and 10 epochs in direction 2. For all hyperparameters we follow [9] and for new proposed losses we scale them down to an order of magnitude corresponding to the basic loss. This setting ensures convergence on the MS-CMRSeg dataset and uncertainty for the two remainings.

References

[1] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

[2] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[3] C. Chen, Q. Dou, H. Chen, J. Qin, and P. A. Heng, “Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2494–2505, Jul. 2020.

[4] F. Wu and X. Zhuang, “Cf distance: A new domain discrepancy metric and application to explicit domain adaptation for cross-modality cardiac image segmentation,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 4274–4285, Dec. 2020.

[5] C. Chen, Q. Dou, Y. Jin, Q. Liu, and P. A. Heng, “Learning with privileged multimodal knowledge for unimodal segmentation,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 3, pp. 621–632, 2021.

[6] C. Lu, S. Zheng, and G. Gupta, “Unsupervised domain adaptation for cardiac segmentation: Towards structure mutual information maximization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2588–2597.

[7] J. Xu and X. Zhuang, “Few-shot learning for cardiac segmentation via self-supervised multi-task learning,” 2021.

[8] C. Pei, F. Wu, L. Huang, and X. Zhuang, “Disentangle domain features for cross-modality cardiac image segmentation,” *Medical Image Analysis*, vol. 71, p. 102078, Jul. 2021.

[9] F. Wu and X. Zhuang, “Unsupervised domain adaptation with variational approximation for cardiac segmentation,” *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3555–3567, 2021.

[10] S. Zhai, R. Gu, W. Lei, and G. Wang, “Myocardial edema and scar segmentation using a coarse-to-fine framework with weighted ensemble,” in *Myocardial pathology segmentation combining multi-sequence CMR challenge*. Springer, 2020, pp. 49–59.

[1] http://www.sdspeople.fudan.edu.cn/zhuangxiahai/0/mscmrseg19/

[2] https://zmiclab.github.io/zxh/0/myops20/

[3] https://github.com/HiLab-git/MyoPS2020

[4] https://github.com/FupingWu90/CT_MR_2D_Dataset_DA