CONTRASTIVE TRANSLATION LEARNING FOR MEDICAL IMAGE SEGMENTATION

*Wankang Zeng, Wenkang Fan, Dongfang Shen, Yinran Chen, Xiongbiao Luo**

Department of Computer Science, Xiamen University, Xiamen 361005, China

ABSTRACT

Unsupervised domain adaptation commonly uses cycle generative networks to produce synthesis data from source to target domains. Unfortunately, translated samples cannot effectively preserve semantic information from input sources, resulting in bad or low adaptability of the network to segment target data. This work proposes an advantageous domain translation mechanism to improve the perceptual ability of the network for accurate unlabeled target data segmentation. Our domain translation employs patchwise contrastive learning to improve the semantic content consistency between input and translated images. Our approach was applied to unsupervised domain adaptation based abdominal organ segmentation. The experimental results demonstrate the effectiveness of our framework that outperforms other methods.

***Index Terms*—** Unsupervised domain adaptation, image segmentation, domain translation, contrastive learning

1. INTRODUCTION

Semantic segmentation is to assign label to each pixel on images. Convolution neural networks (CNN) driven semantic segmentation methods have shown significant improvement in various medical image segmentation tasks. These methods usually require massive labeled data to learn hyperparameter distribution, and also have other limitations. First, pixel-wise data annotation is time-consuming and extremely high cost, especially expert knowledge is entailed in medical issues. Second, the model performance may be deteriorated due to domain shift between training (source) and testing (target) data, e.g., computed tomography (CT) data trained CNN may inaccurately perceive the same semantic class of organs from magnetic resonance (MR) to CT data. This is because that MR and CT are different modalities.

Unsupervised domain adaptation (UDA) for semantic segmentation is promising to resolve these issues [1]. UDA aims to improve the performance of segmentation network on target domain by reducing domain gap between source and target domains. For UDA-driven semantic segmentation, we are interested in training a segmentation network using labeled samples from the source domain and unsupervised target domain samples to accurately predict the region of interest from target domain data. Current UDA methods have been focused on the input space, feature space and output space of CNN [1]. At the input space, the segmentation network takes the translated images generated from unpaired image-to-image translation as the input [2, 3, 4]. The translated images preserve the semantic content of source domain data using cycle consistency loss, and they translate the appearance of target domain to themselves by adversarial learning. At the feature space, some papers aim to learn domain-invariant features for downstream tasks by applying adversarial learning [5], or feature cluster [6]. At the output space, Tsai et al. [7] considered the segmentation result of source and target domain samples share highly semantic structure similarities and employed an adversarial learning scheme on output space to get the prediction for target domain data. Other strategies including self-training [8], entropy minimization [9] and classifier discrepancy [10] are also effective to unsupervised domain adaptation based semantic segmentation.

This work focuses mainly on the input space of CNN to extract rich structural information. Previous approaches [11, 2, 4] applied cycle consistency loss or L_1 reconstruction loss to guarantee that the content of the source input image is preserved in the translation process. However, this pixel-level constraint is difficult to ensure the content consistency in the corresponding region (patch) between the translated image and the original image. If the translated image does not retain the content of the original image, it will result in the segmentation network not being able to effectively perceive the target region. Our image translation driven segmentation framework consists of translation and segmentation models. The translation model uses the patchwise contrast learning to maintain patch content consistency between translated and source images and adversarial learning to migrate the target domain style to translated samples. Furthermore, the segmentation network uses the translated images and the ground-truth source data for parameter learning.

2. METHODS

This section discusses technical details of our methods. Fig. 1 shows our domain translation driven segmentation workflow.

W. Zeng and W. Fan contributed equally. *Corresponding author: xiongbiao.luo@gmail.com. This work was supported partly by the National Natural Science Foundation of China under Grant 61971367 and the Fujian Provincial Technology Innovation Joint Funds under Grant 2019Y9091.

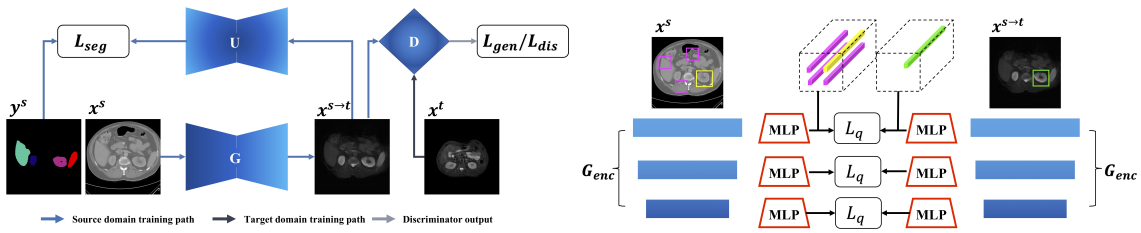


Fig. 1: Our unpaired image translation driven segmentation framework (left) with the patchwise contrastive loss (right). By the supervision of the adversarial losses $\mathcal{L}_{dis}$ and $\mathcal{L}_{gen}$, the generator G produces the translated image $x^{s\to t}$ similar to the target domain, the discriminator D estimates the probability for the distribution of the input data if it is similar to either the translated results or the distribution of target data. The encoder G_{enc} encodes the input into the feature tensor. After generating the outputs of the layer of interest of G_{enc}, MLP embeds the patch randomly selected from feature tensor into the latent space to obtain feature embeddings. The InfoNCE loss $\mathcal{L}_q$ uses these embeddings to measure the similarity between patches.

2.1. Assumptions

Given a set of unpaired source domain data $S = \{x_i^s, y_i^s\}_{i=1}^N$ and target domain data $T = \{x_j^t\}_{j=1}^M$. x^s, x^t denote the input sample. y^s represent the source sample label. We are interested in learning a mapping $G : S \to T$ and obtaining images $\{x^{s\to t} | x^{s\to t} = G(x^s)\}$. Then the data $\{x^{s\to t}, y^s\}$ were applied to train a segmentation network U. The full objective loss $\mathcal{L}_{tot}$ of our network model is formulated as:

$$\mathcal{L}_{tot} = \mathcal{L}_{gen} + \mathcal{L}_{dis} + \mathcal{L}_{seg} + \mathcal{L}_{PatchNCE} \tag{1}$$

where $\mathcal{L}_{gen}$ encourages the generator G to match the distribution of translated images to the data distribution of target domain, and $\mathcal{L}_{dis}$ incentivize the discriminator D to distinguish between translated images $\{x^{s\to t}\}$ and real target images $\{x^t\}$. $\mathcal{L}_{patchNCE}$ is designed to encourage the transformed sample to retain the content of original sample during unpaired image-to-image translation [12]. The segmentation U is constrained by the $\mathcal{L}_{seg}$ to find anatomy of interest from $x^{s\to t}$. The following sections will discuss each loss function.

2.2. Appearance Transformation

Appearance transformation aims to reduce domain gap between source data and target data by using the generator G produce translated image $\{x^{s\to t}\}$ to visually look like to real image in target domain. Such transformation can be achieved by adversarial losses $\mathcal{L}_{gen}$ and $\mathcal{L}_{dis}$. We choose the least squares loss for the adversarial learning and it can mitigate the gradient disappearance problem and stabilize the training process [13]. $\mathcal{L}_{gen}$ and $\mathcal{L}_{dis}$ can be represent by:

$$\mathcal{L}_{gen}(G) = \mathbb{E}_{x^s\sim p(S)}\left[(D(G(x^s)) - 1)^2\right] \tag{2}$$

$$\begin{aligned}\mathcal{L}_{dis}(D) = \mathbb{E}_{x^s\sim p(S)}[D(G(x^s)^2] + \\ \mathbb{E}_{x^t\sim p(T)}[(D(x^t) - 1)^2]\end{aligned} \tag{3}$$

where $p(S)$, $p(T)$ is source domain data distribution, target domain data distribution, respectively.

2.3. Content Consistency

In this section, we firstly review the loss function $\mathcal{L}_{patchNCE}$ in [12]. Specifically, $\mathcal{L}_{patchNCE}$ was employed to maximize mutual information between input samples $\{x^s, x^t\}$and translated samples $\{x^{s\to t}, x^{t\to s}\}$ using the noise contrastive estimation [14, 12]. In other words, we wish for the translated sample to better preserve the content of the input sample at the same spatial location. For example, a patch including right kidney of $x^{s\to t}$ should better maintain the structure correspondence with the patch including right kidney of x^s, less so than the other patches of x^s. $\mathcal{L}_{patchNCE}$ is defined as:

$$\begin{aligned}\mathcal{L}_{patchNCE} = \mathcal{L}^S_{patchNCE}(x^s, x^{s\to t}) + \\ \mathcal{L}^T_{patchNCE}(x^t, x^{t\to s})\end{aligned} \tag{4}$$

$$\mathcal{L}^S_{patchNCE}(x^s, x^{s\to t}) = \mathbb{E}_{x^s\sim S}\sum_{l=1}^{L}\sum_{s=1}^{S_l} L_q(\hat{u}_l^s, u_l^s, u_l^{S_l\setminus s}) \tag{5}$$

$$\{u_l\}_{l=1}^L = \{F_l(G^l_{enc}(x^s))\}_{l=1}^L \tag{6}$$

$$\{\hat{u}_l\}_{l=1}^L = \{F_l(G^l_{enc}(x^{s\to t}))\}_{l=1}^L \tag{7}$$

$$\mathcal{L}^T_{patchNCE}(x^t, x^{t\to s}) = \mathbb{E}_{x^t\sim T}\sum_{l=1}^{L}\sum_{s=1}^{S_l} L_q(\hat{v}_l^s, v_l^s, v_l^{S_l\setminus s}) \tag{8}$$

$$\{v_l\}_{l=1}^L = \{F_l(G^l_{enc}(x^t))\}_{l=1}^L \tag{9}$$

$$\{\hat{v}_l\}_{l=1}^L = \{F_l(G^l_{enc}(x^{t\to s}))\}_{l=1}^L \tag{10}$$

where G_{enc} is the encoder of generator G, F is a multilayer perceptron (MLP) with two hidden layers that embeds encoded feature to latent space where the InfoNCE loss $\mathcal{L}_q$ is calculated [15]. In the context of the encoder G_{enc}, deeper layers have bigger patches. We select the output of L layers of G_{enc} and randomly sample feature vectors from $\{G^l_{enc}(x)\}_{l=1}^L$ to compute the latent representation

Table 1: Dice and ASSD of using the five compared segmentation approaches

Approaches	Dice (STD)					ASSD (STD)				
	Spleen	Left Kidney	Right Kidney	Liver	Mean	Spleen	Left Kidney	Right Kidney	Liver	Mean
M1 [4]	37.9 (10.8)	65.8 (3.2)	59.2 (8.0)	66.5 (4.8)	57.4	7.1 (1.3)	4.5 (1.4)	5.0 (1.5)	3.8 (0.6)	5.0
M2 [2]	49.7 (12.6)	76.9 (3.8)	76.7 (5.9)	77.8 (4.7)	70.3	7.0 (2.2)	4.4 (1.1)	1.9 (1.0)	3.5 (0.6)	4.2
M3 [7]	**84.2 (7.4)**	80.3 (4.6)	84.7 (6.3)	66.2 (4.9)	78.9	**1.4 (1.4)**	3.2 (1.8)	**1.6 (1.8)**	4.3 (0.8)	2.6
M4 [11]	75.1 (8.7)	**88.3 (1.5)**	**85.3** (6.3)	68.1 (4.5)	79.2	3.3 (0.4)	**2.0 (2.2)**	3.0 (1.6)	**1.8 (1.0)**	2.5
Ours	76.6 (10.5)	76.6 (10.5)	83.3 (5.5)	**83.4 (1.2)**	**82.3**	3.5 (2.7)	2.1 (0.4)	2.1 (1.7)	2.4 (0.5)	2.5

$\{u_l, \hat{u}_l, v_l, \hat{v}_l\}_{l=1}^{L}$. S_l denotes the number of feature vectors at the l-th layer of encoder G_{enc}. $\hat{u}_l^s \in \mathbb{R}^{C_l}$ and $u_l^s \in \mathbb{R}^{C_l}$ corresponds to the embeddings at the same spatial location of feature maps $G_{enc}^l(x^{s\to t})$ and $G_{enc}^l(x^s)$. $u_l^{S_l \setminus s} \in \mathbb{R}^{(S_l-1)\times C_l}$ and $v_l^{S_l \setminus s}$ exclude u_l^s, $\hat{v}_l^s$ and v_l^s, respectively. C_l represents the length of the feature embedding. the InfoNCE loss function $\mathcal{L}_q$ is a contrastive loss, which aims to classify the feature embeddings into either similar or dissimilar samples [16], and it can be calculated by [12, 14, 17]:

$$\mathcal{L}_q(q, k^+, k^-) = -log \frac{exp(q \cdot k^+/\tau)}{exp(q \cdot k^+/\tau) + \sum_{n=1}^{\hat{N}} exp(q \cdot k_n^-/\tau)} \tag{11}$$

where symbol $\cdot$ denotes the dot product to measure the similarity between embeddings and τ determines the scale level of the intensity similarity [12, 17]. Although $\mathcal{L}_q$ and the soft nearest neighbor loss function have a similar role, the significant difference is that the former only uses one positive pair and the latter multiple positive pairs. The latter is more suitable for multimodal source domain adaptation scenarios.

2.4. Segmentation Network Training

The perception of the target-data segmentation network can be improved using the sample pairs $\{x^{s\to t}, y^s\}$ as the input since $\{x^{s\to t}\}$ carries an enhanced visual appearance resemblance to the target ones compared to source samples $\{x^s\}$.The segmentation loss function $\mathcal{L}_{seg}$ is calculated by

$$\mathcal{L}_{seg}(y^{s\to t}, y^s) = \mathcal{L}_{wce}(y^{s\to t}, y^s) + \mathcal{L}_{dl}(y^{s\to t}, y^s) \tag{12}$$

$$\mathcal{L}_{wce}(y^{s\to t}, y^s) = -\sum_{i=1}^{N}\sum_{c=1}^{N_C} w_c y_i^{c_s} log y_i^{c_{s\to t}} \tag{13}$$

$$\mathcal{L}_{dl}(y^{s\to t}, y^s) = 1 - \sum_{c=1}^{N_C} \frac{\sum_{i=1}^{N} y_i^{c_{s\to t}} y_i^{c_s}}{\sum_{i=1}^{N}(y_i^{c_{s\to t}})^2 + \sum_{i=1}^{N}(y_i^{c_s})^2} \tag{14}$$

The loss function $\mathcal{L}_{seg}$ is made up of the weight cross entropy $\mathcal{L}_{wce}$ and the dice loss $\mathcal{L}_{dl}$. $y^{s\to t} = U(x^{s\to t})$ is the prediction result of $x^{s\to t}$. $y_i^{c_{s\to t}}$ represents the probability of pixel i belongs to class C, and $y_i^{c_s}$ indicates the ground truth label for pixel i of y_i^s. N_C is the number of classes, and $w_c = 1 - \frac{\sum_{i=1}^{N} y_i^c}{N}$ denotes the weight of class c. The hybrid loss $\mathcal{L}_{seg}$ can alleviate the class imbalance problem in training [11].

3. EXPERIMENTAL SETTINGS

We test our proposed method on the cross-modality adaptation to segment four abdominal organs from CT to MR images. We used 20 samples of MR data from the CHAOS Challenge [18] and 30 samples of CT data from [19]. We follow the ratio in the [11] to split with 20% for testing and 80% samples for training. For data preprocessing, we truncated the intensity of CT images to [-200, 300] to accentuate the abdominal organs details, followed by were normalized to [0,1]. For MR scans, the intensities were normalized to [0, 1] and it same as the preprocessing in [2]. Furthermore, we utilized axial view slices of volumetric data and resampled their size to 256×256. We also employed some data augmentation strategies to prevent different approaches to over-fitting, such as random sharpness transformation, random brightness contrast transformation and elastic deformations.

The structure of our generator and discriminator are the same as [2, 11]. We employed the 2D U-Net-like neural network [3] as the segmentation network. These architecture design has been proved sound and effective according to their experimental results [2, 11, 3, 4]. Our network parameters were initialized using the previous method [17]. Two Adam optimizers were deployed to update parameters. One optimize the generator and the segmentation network, another optimize the discriminator. Their initial learning rate are 0.0002.

4. RESULTS AND DISCUSSION

We evaluate and compare our approach to M1 [4], M2 [2], M3 [7], and M4 [11]. Our evaluation metrics are dice coefficient (Dice) and average symmetric surface distance (ASSD) used to quantitatively compare the performance of different methods. Table 1 quantifies the segmentation result of the compared method for abdominal MR images. The average Dice of the segmentation result of M1, M2, M3, M4, and our method was 57.4, 70.3, 78.9, 79.2, and 82.3, respectively, while the average ASSD of M1, M2, M3, M4, and our approach was 5.0, 4.2, 2.6, 2.5, and 2.5. While the ASSD of M4 and our method was comparable, the Dice of our approach was much better than the other four methods.

Fig. 2 illustrates translated image generated from M1 [4], M2 [2], M3 [7], M4 [11], and our method for the same inputs. The qualitative results of translated results demonstrate

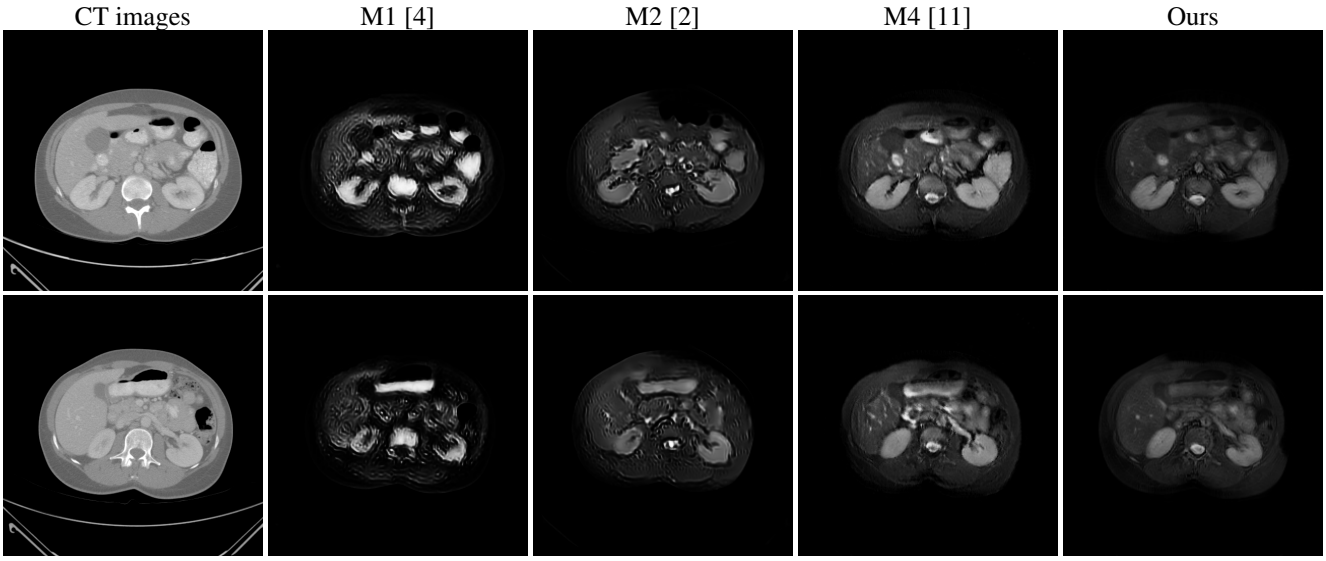
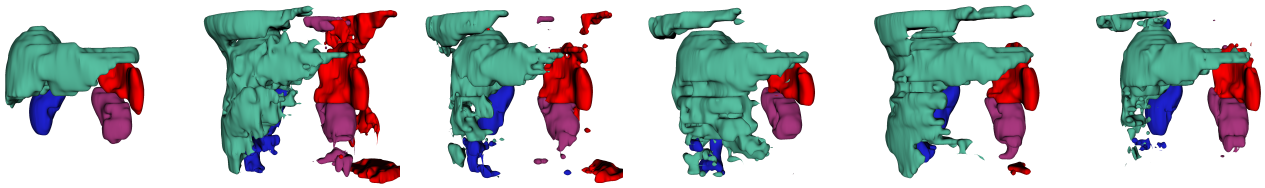


Fig. 2: Visual comparison of CT-to-MR translated images produced by the different translation methods



(a) Ground Truth (b) M1 [4] (c) M2 [2] (d) M3 [7] (e) M4 [11] (f) Ours

Fig. 3: Qualitative segmentation results of using the five different segmentation approaches. The segmented liver, spleen, left kidney, and right kidney regions are displayed in various colors of cyan, red, pink, and blue, respectively.

M4 and our method can better preserve the semantic content of the input CT image. The translated images of M1 and M2 have obvious change in structure and content. M3 is an output space domain adaptation method and it does not perform translation transformations on input samples. Furthermore, Fig. 3 visually compares the segmentation results of using different approaches M1, M2, M3, M4 and our approach. Although the five methods have false positive prediction, the liver segmentation result of our method visually looks much better than the other four compared methods.

This work aims to use the translated results from labeled source domain data to better adapt segmentation network to unlabeled target samples. We proposed a new image translation driven segmentation method. The experimental results demonstrate our method achieve the comparable or better results than other methods in terms of the qualitative and quantitative assessment. However, we have not quantitatively analyzed the effect of the quality of translated images on the segmentation network. We only demonstrate the validity of the translated images in terms of the average Dice and ASSD of the segmentation results. Therefore, a new evaluation metric that can evaluate the semantic content consistency between translated images and the input and measure the distribution difference between translated images and target one will be investigated in the future. On the other hand, our method still has many false positive or negative prediction. Integrating other domain adaptation strategies into our approach can further improve the segmentation. Additionally, we need to further validate the performance of our proposed method on different types of domain adaptation problems.

5. CONCLUSION

This work proposes an image domain translation driven segmentation framework that integrates patchwise contrastive learning based unpaired images into unsupervised adaptation and uses a 2D U-Net-like segmentation network to extract abdominal organs of interest from unlabeled MR images. We evaluate our approach in the CT-to-MR cross-modal segmentation task of abdominal organs. The experimental results demonstrate the effectiveness of our proposed method in adapting the segmentation network to unlabeled MR data.

6. REFERENCES

[1] Marco Toldo, Andrea Maracani, Umberto Michieli, and Pietro Zanuttigh, "Unsupervised domain adaptation in semantic segmentation: a review," *Technologies*, vol. 8, no. 2, pp. 35, 2020.

[2] Yuankai Huo, Zhoubing Xu, Hyeonsoo Moon, Shunxing Bao, Albert Assad, Tamara K Moyo, Michael R Savona, Richard G Abramson, and Bennett A Landman, "Synseg-net: Synthetic segmentation without target modality ground truth," *IEEE transactions on medical imaging*, vol. 38, no. 4, pp. 1016–1025, 2018.

[3] Wankang Zeng, Wenkang Fan, Rong Chen, Zhuohui Zheng, Song Zheng, Jianhui Chen, Rong Liu, Qiang Zeng, Zengqin Liu, Yinran Chen, et al., "Accurate 3d kidney segmentation using unsupervised domain translation and adversarial networks," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 598–602.

[4] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[5] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. PMLR, 2018, pp. 1989–1998.

[6] Marco Toldo, Umberto Michieli, and Pietro Zanuttigh, "Unsupervised domain adaptation in semantic segmentation via orthogonal and clustered embeddings," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1358–1368.

[7] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7472–7481.

[8] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang, "Confidence regularized self-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5982–5991.

[9] Minghao Chen, Hongyang Xue, and Deng Cai, "Domain adaptation for semantic segmentation with maximum squares loss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2090–2099.

[10] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3723–3732.

[11] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng Ann 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, 2020.

[12] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu, "Contrastive learning for unpaired image-to-image translation," in *European Conference on Computer Vision*. Springer, 2020, pp. 319–345.

[13] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[14] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[16] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *arXiv preprint arXiv:2006.10511*, 2020.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[18] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al., "Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation," *Medical Image Analysis*, vol. 69, pp. 101950, 2021.

[19] B Landman, Z Xu, J Eugenio Igelsias, M Styner, T Langerak, and A Klein, "Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge," in *Proc. MICCAI: Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge*, 2015.