MDT-NET: MULTI-DOMAIN TRANSFER BY PERCEPTUAL SUPERVISION FOR UNPAIRED IMAGES IN OCT SCAN

Weinan Song† Gaurav Fotedar† Nima Tajbakhsh† Ziheng Zhou⋆ Lei He⋆ Xiaowei Ding‡

⋆ University of California, Los Angeles
† VoxelCloud Inc
‡ Shanghai Jiaotong University

ABSTRACT

Deep learning models tend to underperform in the presence of domain shifts. Domain transfer has recently emerged as a promising approach wherein images exhibiting a domain shift are transformed into other domains for augmentation or adaptation. However, with the absence of paired and annotated images, models merely learned by adversarial loss and cycle consistency loss could result in poor consistency of anatomy structures during the translation. Additionally, the complexity of learning multi-domain transfer could significantly increase with the number of target domains and source images. In this paper, we propose a multi-domain transfer network, named MDT-Net, to address the limitations above through perceptual supervision. Specifically, our model consists of a single encoder-decoder network and multiple domain-specific transfer modules to disentangle feature representations of the anatomy content and domain variance. Owing to this architecture, the model could significantly reduce the complexity when the translation is conducted among multiple domains. To demonstrate the performance of our method, we evaluate our model qualitatively and quantitatively on RETOUCH, an OCT dataset comprising scans from three different scanner devices (domains). Furthermore, we take the transfer results as additional training data for fluid segmentation to prove the advantage of our model indirectly, i.e., in the task of data adaptation and augmentation. Experimental results show that our method could bring universal improvement in these segmentation tasks, which demonstrates the effectiveness and efficiency of MDT-Net in multi-domain transfer.

Index Terms—Domain Transfer, Data Augmentation, OCT Segmentation

1. INTRODUCTION

Deep learning has proved effective in automating the diagnosis and quantification of various diseases and conditions. These models, however, tend to underperform in the presence of domain shifts, which commonly exist in medical imaging due to variance of scanner devices [1], diversity of imaging protocols [2], or deviation between real and synthesized data [3]. In the absence of paired and labeled data, models based on cycle-consistent loss [4][5] could achieve the image-to-image translation by simply learning from a cycle transfer process within two domains. However, such models are easy to lose the content consistency on diseased images (as shown in our experiment) during the transfer. Besides, the model can only learn a one-to-one domain transfer at one time, where the model complexity grows geometrically with the number of domains. In comparison, neural style transfer [6] provides a promising solution to keep the content consistency by aligning the statistical distribution of medical images collected from different sources [7][8][9], while the model complexity problem remains as the optimization time increases linearly with the number of images in source domains.
To address the limitations above, we introduce MDT-Net, a multi-domain transfer model, to decouple the feature representation of anatomy structures and domain deviation of medical images with an encoder-decoder network and multiple domain transfer modules. Inspired by perceptual supervision [6], MDT-Net preserves anatomical structures during translation by imposing content loss during the identical domain transfer and domain loss (some may call style loss) in the diverse domain transfer. Therefore, it can directly translate images into multiple target domains at one time without any reference images during the inference, where the translation time is independent of the number of source images. Moreover, the model complexity reduces from $\frac{n(n-1)}{2}$ to $n$ when involving transfer among $n$ domains compared with cycle-consistent-based models. To demonstrate the translation performance, we first compare the translated results in domain variance against the target domains and content similarity with the source content. Then we take these translated images as extra data to boost existing segmentation models. Extensive results show that our model can significantly outperform other methods qualitatively and quantitatively.

2. METHODOLOGIES

2.1. Keeping Anatomy Consistency during Translation

As shown in Figure 1, MDT-Net consists of an encoder-decoder network ($f_e()$ and $f_d()$) to learn anatomy-consistent feature representation and multiple feature transfer modules ($t_i()$, $i = 1, 2, \ldots, X$) to learn domain transition, where each module learns feature translation to a target domain. The training process during the translation is composed of two circumstances: 1) identical domain transfer, where the model generates an image $I'$ by $f_d(f_e(I))$ from an image $I$ in the source domain, and 2) diverse domain transfer, where the model outputs a translated image $I_X'$ into the target domain $X$ via $f_d(t_X(f_e(I)))$. Since the domain transfer toward each target domain is learned explicitly by a feature transfer module, MDT-Net can directly translate images into multiple target domains without any reference images by forwarding deep features into different feature transfer modules respectively during the inference.

2.2. Perceptual Supervision

Perceptual supervision is first proposed in [6] and has been widely applied in style transfer between paintings and photographs by capturing implicit content features and texture statistics. Generally, the perceptual loss is calculated by a pre-trained feature extraction network (FEN) $F$ to compute the reconstruction loss over content and style features. In our model, we define the loss function as a combination of $L_{content}$ and $L_{domain}$ following the perceptual loss as:

$$L = L_{content}(I, I') + \sum_X \alpha_X \cdot L_{domain}^X(I_X, I'_X),$$  

where $I_X$ is the image randomly sampled in a target domain $X$. The content loss $L_{content}$ and domain (style) loss $L_{domain}$ are defined as:

$$L_{content} = \frac{1}{N_c} \sum_l \| F^l(I') - F^l(I) \|^2$$

$$L_{domain} = \frac{1}{N_d} \sum_l \| G(F^l(I_X')) - G(F^l(I_X)) \|^2.$$  

$F^l(\cdot)$ denotes the features selected from the FEN and $G(\cdot)$ is a function to compute the Gram matrix [6], which has been used to compare the texture statistics in paintings.

2.3. Network Architecture

Our network architecture is developed based on StyleBank [11]. We make several improvements to accommodate it to the domain transfer problem in medical images: 1) We apply transfer modules on multi-level features generated by the encoding network to learn feature translation. 2) The transfer modules of MDT-Net consist of multiple dense-connected convolution layers instead of a single convolution layer. 3) The model is trained to predict a residual image instead of the transfer result directly. Evaluation of these changes and generation comparisons between StyleBank and MDT-Net can be seen in the ablation study in section 4.

3. EXPERIMENT

3.1. Dataset

We use RETOUCH [12] to validate the domain transfer capability of MDT-Net. The dataset contains 70 OCT scans taken by three different vendors (domains): 1) 24 from Cirrus, 2) 24 from Spectralis, and 3) 22 from Topcon. Each image is annotated with three kinds of pathological annotations i.e., intraretinal fluid, subretinal fluid and pigment epithelial detachment for segmentation. We use $C, S, T$ to represent each domain and randomly select 5 cases from each domain as test data for both domain transfer and segmentation tasks. To be noted, annotations are only used in segmentation models.

3.2. Evaluation

We use Fréchet Inception Distance (FID) [13] and Learned Perceptual Image Patch Similarity (LPIPS) [14] to compare the domain similarity and content inconsistency of generated images. We also propose Domain Perceptual Distance (DPD), a combination of FID and LPIPS, as an overall evaluation metric to indicate the distance to the optimal results by:

$$DPD = FID + \lambda \cdot (1 - LPIPS) \times 100\%,$$  

where we set $\lambda = 1$ in this paper. For data adaptation and augmentation, we use averaged dice scores of the three segmentation targets as the evaluation metric. For comparison,
we train four other unsupervised domain transfer models that are either based on cycle-consistent learning, i.e., CycleGAN [4] and StarGAN2 [15], or perceptual learning, i.e., AdaIN [16] and StyleBank.

3.3. Training

Due to the various number of slices in three domains, we train the model for 32, 80, and 40 epochs for $C$, $S$, and $T$. The learning rate starts from $10^{-3}$ and decays by 0.1 in the middle. We use $\alpha_X = 10$ in $L_{percep}$ and select the same layers as in [6] from VGG-16 [17] to obtain $L_{content}$ and $L_{domain}$.

4. RESULTS

4.1. Multi-domain Transfer

We first directly compare the translated images under the transfer among the three domains and show the results in Figure 3 and Table 1. Our proposed method can achieve the best balance between domain shift and content consistency. Compared with MDT-Net, CycleGAN can get excellent performance on domain shift but fails to keep the anatomical structure of the retina. StyleBank preserves the content during transformation but can not reasonably match the textures of target domains. As StarGAN2 fails to retain the content, we exclude its results in the latter experiments.

Table 1. Quantitative comparison of transfer results averaged in six types of domain transfer by FID, LPIPS and DPD.

| Method        | CycleGAN | StarGAN2 | AdaIN   | StyleBank | Ours        |
|---------------|----------|----------|---------|-----------|-------------|
| FID           | 47.62    | 276.45   | 126.51  | 129.47    | 56.23       |
| LPIPS         | 62.72    | 43.48    | 53.88   | 86.85     | 75.67       |
| DPD           | 84.91    | 332.96   | 172.62  | 142.62    | 80.56       |

4.2. Data Adaptation

In this experiment, we assume that only images in one domain are provided with annotations, while the segmentation model is expected to accommodate to the test data in the other two domains. This is a very common situation in clinical applications where the inference data could be collected from other sources (domains). We use DeeplabV3+ [18] as the baseline segmentation model. We first train the model with images from one domain, then add translated images that share the same annotations with the source images. Therefore, the improved performance brought by these additional images can indicate the quality of domain transfer results. For example, for domain $C$, the gap between dice scores of the models trained with 1) $C$ only and 2) $C \rightarrow S$, and $C \rightarrow T$ can indicate the improved adaptation ability in unseen domains, i.e., $S$ and $T$. As shown in Table 2, MDT-Net brings the biggest improvement and demonstrate the best transfer results.
Table 2. Evaluating translated images by DeeplabV3+ in model adaptation.

| Method   | C       | S       | T       | Avg   |
|----------|---------|---------|---------|-------|
|          | S       | T       | C       |       |
| -        | 59.5    | 70.0    | 59.1    | 64.8  |
| CycleGAN | 53.1    | 55.9    | 52.2    | 60.8  |
| AdaIN    | 67.2    | 77.6    | 56.4    | 65.9  |
| StyleBank| 76.0    | 83.4    | 62.5    | 84.4  |
| Ours     | 84.5    | 76.9    | 67.5    | 83.0  |

Table 3. Evaluating translated images by multiple segmentation models in dice score for data augmentation.

| Seg Model | Method   | C     | S     | T     | Avg   |
|-----------|----------|-------|-------|-------|-------|
| U-Net     |          |       |       |       |       |
|           | CycleGAN | 76.64 | 78.36 | 67.02 | 74.01 |
|           | AdaIN    | 68.24 | 74.71 | 77.64 | 73.53 |
|           | StyleBank| 72.05 | 74.40 | 79.77 | 75.41 |
|           | Ours     | 83.97 | 71.74 | 73.35 | 76.35 |
| Deeplab   |          |       |       |       |       |
|           | CycleGAN | 79.73 | 88.90 | 85.53 | 84.72 |
|           | AdaIN    | 80.66 | 87.59 | 85.86 | 84.70 |
|           | StyleBank| 83.44 | 88.83 | 80.52 | 84.13 |
|           | Ours     | 81.61 | 89.16 | 87.63 | 86.13 |

4.3. Data Augmentation

Unlike the adaptation task where images from other domains are unseen during training, we take all images, including both original images from three domains and the transfer results generated by the six kinds of domain transfer within these three domains, as the training data for segmentation in this experiment. Similarly, we use improvements in segmentation accuracy to indicate the domain transfer performance. To avoid influence brought by variation in segmentation models, we introduce U-Net [19] and HR-Net [20] as additional baseline models. As shown in Table 3, our method can still best boost all three existing segmentation models, bringing about +2% in dice scores.

4.4. Ablation Study

In this experiment, we change the architecture of MDT-Net for ablation study. MDT-Net_D removes the residual learning at the output, where the decoder directly generates the translated result. MDT-Net_S replaces the feature transfer module with the style bank structure as proposed in [11]. MDT-Net_J9 replaces VGG-16 with VGG-19 as FEN. From Table 4 we can find that our proposed feature transfer modules play the most important role during the domain transfer. Combined with Figure 3, we can see that changing FEN does not affect the result while removing the residual structure mainly increases the converging time.

Table 4. Evaluation of translated images in the ablation study.

| Method   | MDT-Net_D | MDT-Net_S | MDT-Net_J9 | Baseline |
|----------|-----------|-----------|------------|----------|
| FIDₜ     | 60.95     | 125.46    | 57.45      | 56.23    |
| LPIPS    | 76.05     | 67.63     | 74.98      | 75.67    |
| DPDₜ     | 84.90     | 157.83    | 82.47      | 80.56    |

5. CONCLUSION

In this paper, we introduce MDT-Net to achieve multi-domain transfer within one single model trained by unpaired and unlabeled images with perceptual supervision. We disentangle the anatomy content and domain variance by an encoder-decoder network and multiple domain-specific transfer modules. Furthermore, extensive experiments on the task of transfer among three domains of OCT images have validated the advantage of MDT-Net qualitatively and quantitatively.

6. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by [12]. Ethical approval was not required as confirmed by the license attached with the open access data.

7. ACKNOWLEDGMENTS

No funding was received for this study. The authors have no relevant financial or non-financial interests to disclose.

8. REFERENCES

[1] Shahab Aslani, Vittorio Murino, Michael Dayan, Roger Tam, Diego Sona, and Ghassan Hamarneh, “Scanner invariant multiple sclerosis lesion segmentation from mri,” in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, 2020, pp. 781–785.

[2] Mohsen Ghafoorian, Alireza Mehtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik de Leeuw, Clare M
Tempany, Bram Van Ginneken, et al., “Transfer learning for domain adaptation in mri: Application in brain lesion segmentation,” in International conference on medical image computing and computer-assisted intervention. Springer, 2017, pp. 516–524.

[3] Weinan Song, Yuan Liang, Jiawei Yang, Kun Wang, and Lei He, “Oral-3d: Reconstructing the 3d structure of oral cavity from panoramic x-ray,” in Proceedings of the AAAI Conference on Artificial Intelligence, 2021, vol. 35, pp. 566–573.

[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] Mengwei Ren, Neel Dey, James Fishbaugh, and Guido Gerig, “Segmentation-renormalized deep feature modulation for unpaired image harmonization,” IEEE Transactions on Medical Imaging, vol. 40, no. 6, pp. 1519–1530, 2021.

[6] Leon A Gatys, Alexander S Ecker, and Matthias Bethge, “Image style transfer using convolutional neural networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2414–2423.

[7] Chunwei Ma, Zhanghexuan Ji, and Mingchen Gao, “Neural style transfer improves 3d cardiovascular mr image segmentation on inconsistent data,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2019, pp. 128–136.

[8] Pietro Antonio Cicalesre, Aryan Mobiny, Pengyu Yuan, Jan Becker, Chandra Mohan, and Hien Van Nguyen, “Stypath: Style-transfer data augmentation for robust histology image classification,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2020, pp. 351–361.

[9] Agnieszka Mikolajczyk and Michal Grochowski, “Style transfer-based image synthesis as an efficient regularization technique in deep learning,” in 2019 24th International Conference on Methods and Models in Automation and Robotics (MMAR). IEEE, 2019, pp. 42–47.

[10] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in European conference on computer vision. Springer, 2016, pp. 694–711.

[11] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua, “Stylebank: An explicit representation for neural image style transfer,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1897–1906.

[12] Hrvoje Bogunović, Freerk Venhuizen, Sophie Klimscha, Stefanos Apostolopoulos, Alireza Bab-Hadiashar, Ulas Bagci, Mirza Faisal Beg, Loza Bekalo, Qiang Chen, Carlos Ciller, et al., “Retouch: the retinal oct fluid detection and segmentation benchmark and challenge,” IEEE transactions on medical imaging, vol. 38, no. 8, pp. 1858–1874, 2019.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” Advances in neural information processing systems, vol. 30, 2017.

[14] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.

[15] Yunjey Choi, Youngjun Uh, Jaejun Yoo, and Jung-Woo Ha, “Stargan v2: Diverse image synthesis for multiple domains,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8188–8197.

[16] Xun Huang and Serge Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1501–1510.

[17] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[18] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas 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.

[20] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaozhi Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al., “Deep high-resolution representation learning for visual recognition,” IEEE transactions on pattern analysis and machine intelligence, 2020.