DiMix: Disentangle-and-Mix Based Domain Generalizable Medical Image Segmentation

Hyeongyu Kim[1], Yejee Shin[1,2], and Dosik Hwang[1,3,4,5(✉)]

[1] School of Electrical and Electronic Engineering, Yonsei University, Seoul, Republic of Korea
{lion4309,yejeeshin,dosik.hwang}@yonsei.ac.kr
[2] Probe Medical, Seoul, Republic of Korea
[3] Department of Radiology and Center for Clinical Imaging Data Science, Yonsei University, Seoul, Republic of Korea
[4] Department of Oral and Maxillofacial Radiology, College of Dentistry, Yonsei University, Seoul, Republic of Korea
[5] Center for Healthcare Robotics, Korea Institute of Science and Technology, Seoul, Republic of Korea

Abstract. The rapid advancements in deep learning have revolutionized multiple domains, yet the significant challenge lies in effectively applying this technology to novel and unfamiliar environments, particularly in specialized and costly fields like medicine. Recent deep learning research has therefore focused on domain generalization, aiming to train models that can perform well on datasets from unseen environments. This paper introduces a novel framework that enhances generalizability by leveraging transformer-based disentanglement learning and style mixing. Our framework identifies features that are invariant across different domains. Through a combination of content-style disentanglement and image synthesis, the proposed method effectively learns to distinguish domain-agnostic features, resulting in improved performance when applied to unseen target domains. To validate the effectiveness of the framework, experiments were conducted on a publicly available Fundus dataset, and comparative analyses were performed against other existing approaches. The results demonstrated the power and efficacy of the proposed framework, showcasing its ability to enhance domain generalization performance.

Keywords: Domain generalization · Medical image segmentation · Disentanglement · Transformers

1 Introduction

Deep learning has achieved remarkable success in various computer vision tasks, such as image generation, translation, and semantic segmentation [3,7,8,16,23]. However, a limitation of deep learning models is their restricted applicability

H. Kim and Y. Shin—Equal contribution.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 242–251, 2023.
https://doi.org/10.1007/978-3-031-43898-1_24

to the specific domains they were trained on. Consequently, these models often struggle to generalize to new and unseen domains. This lack of generalization capability can result in decreased performance and reduced applicability of models, particularly in fields such as medical imaging where data distribution can vary greatly across different domains and institutions [14,18].

Methods such as domain adaptation (DA) or domain generalization (DG) have been explored to address the aforementioned problems. These methods aim to leverage learning from domains where information such as annotations exists and apply it to domains where such information is absent. Unsupervised domain adaptation (UDA) aims to solve this problem by simultaneously utilizing learning from a source domain with annotations and a target domain without supervised knowledge. UDA methods are designed to mitigate the issue of domain shift between the source and target domains. Pixel-level approaches, as proposed in [2,4,9,17], focus on adapting the source and target domains at the image level. These UDA methods, based on image-to-image translation, effectively augment the target domain when there is limited domain data. Manipulating pixel spaces is desirable as it generates images that can be utilized beyond specific tasks and easily applied to other applications.

In DG, unlike UDA, the model aims to directly generalize to a target domain without joint training or retraining. DG has been extensively studied recently, resulting in various proposed approaches to achieve generalization across domains. One common approach [6] is adversarial training, where a model is trained to be robust to domain shift by incorporating a domain discriminator into the training process. Another popular approach [1,13] involves using domain-invariant features, training the model to learn features not specific to any particular domain but are generalizable across all domains.

DG in the medical field includes variations in imaging devices, protocols, clinical centers, and patient populations. Medical generalization is becoming increasingly important as the use of medical imaging data is growing rapidly. Compared to general fields, DG in medical fields is still in its early stages and faces many challenges. One major challenge is the limited amount of annotated data available. Additionally, medical imaging data vary significantly across domains, making it difficult to develop models that generalize well to unseen domains. Recently, researchers have made significant progress in developing domain generalization methods for medical image segmentation. [10] learns a representative feature space through variational encoding with a linear dependency regularization term, capturing the shareable information among medical data collected from different domains. Based on data augmentation, [5,11,21] aims to solve domain shift problems with different distributions. [5], for instance, proposes utilizing domain-discriminative information and content-aware controller to establish the relationship between multiple source domains and an unseen domain. Based on alignment learning, [18] introduces enhancing the discriminative power of semantic features by augmenting them with domain-specific knowledge extracted from multiple source domains. Nevertheless, there exists an opportunity for further enhancement in effectively identifying invariant fea-

tures across diverse domains. While current methods have demonstrated notable progress, there is still scope for further advancements to enhance the practical applicability of domain generalization in the medical domain.

In recent years, Transformer has gained significant attention in computer vision. Unlike traditional convolutional neural networks (CNNs), which operate locally and hierarchically, transformers utilize a self-attention mechanism to weigh the importance of each element in the input sequence based on its relationship with other elements. Swin Transformer [12] has gained significant attention in computer vision due to its ability to capture global information effectively in an input image or sequence. The capability has been utilized in disentanglement-based methods to extract variant features (e.g., styles) for synthesizing images. [20] has successfully introduced Swin-transformers for disentanglement into StyleGAN modules [8], leading to the generation of high-quality and high-resolution images. The integration of Transformer models into the medical domain holds great promise for addressing the challenges of boosting the performance of domain generalization.

In this paper, we present a novel approach for domain generalization in medical image segmentation that addresses the limitations of existing methods. To be specific, our method is based on the disentanglement training strategy to learn invariant features across different domains. We first propose a combination of recent vision transformer architectures and style-based generators. Our proposed method employs a hierarchical combination strategy to learn global and local information simultaneously. Furthermore, we introduce domain-invariant representations by swapping domain-specific features, facilitating the disentanglement of content (e.g., objects) and styles. By incorporating a patch-wise discriminator, our method effectively separates domain-related features from entangled ones, thereby improving the overall performance and interpretability of the model. Our model effectively disentangles both domain-invariant features and domain-specific features separately. Our proposed method is evaluated on a medical image segmentation task, namely retinal fundus image segmentation with four different clinical centers. It achieves superior performance compared to state-of-the-art methods, demonstrating its effectiveness.

2 Methods

2.1 Framework

Efficiently extracting information from input images is crucial for successful domain generalization, and the process of reconstructing images using this information is also important, as it allows meaningful information to be extracted and, in combination with the learning methods presented later, allows learning to discriminate between domain-relevant and domain-irrelevant information. To this end, we designed an encoder using a transformer structure, which is nowadays widely used in computer vision, and combined it with a StyleGAN-based image decoder.

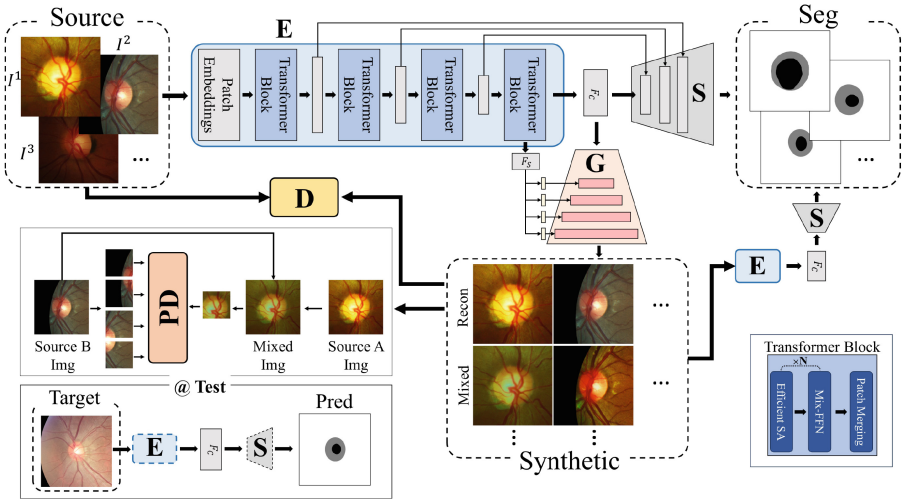

Fig. 1. Overview of our proposed framework.

The overall framework of our approach comprises three primary architectures: an encoder denoted as E, a segmentor denoted as S, and an image generator denoted as G. Additionally, the framework includes two discriminators: D and PD, as shown in Fig. 1.

The encoder E is constructed using hierarchical transformer architecture, which offers increased efficiency and flexibility through its consideration of multi-scale features, as documented in [12,19,20]. Hierarchical transformers enable efficient extraction of diverse data representations and have computational advantages over conventional vision transformers. In our proposed method, the transformer encoder consists of three key mechanisms: Efficient SA, Mix-FFN, and Patch Merging. Each of these mechanisms plays a significant role in capturing and processing information within the transformer blocks. The Efficient SA mechanism involves computing the query (Q), key (K), and value (V) heads using the self-attention mechanism. To reduce the computational complexity of the attention process, we choose a reduction ratio (R), which allows us to decrease the computational cost. K is reshaped using the operation Reshape$(\frac{N}{R}, C \cdot R)(K)$, where N represents the number of patches and C denotes the channel dimension. In the Mix-FFN mechanism, we incorporate a convolutional layer in the feed-forward network (FFN) to consider the leakage of location information. The process is expressed as:

$$F_{out} = \text{MLP}(\text{GELU}(\text{Conv}(\text{MLP}(F_{in})))) + F_{in}, \tag{1}$$

where F_{in} and F_{out} represent the input and output features, respectively. This formulation enables the model to capture local continuity while preserving important information within the feature representation. To ensure the preservation of local continuity across overlapping feature patches, we employ the Patch

Merging process. This process combines feature patches by considering patch size (K), stride (S), and padding size (P). For instance, we design the parameters as $K = 7, S = 4, P = 3$, which govern the characteristics of the patch merging operation.

E takes images $I^{\mathcal{D}}$ from multiple domains $\mathcal{D} : \{\mathcal{D}_1, \mathcal{D}_2, ...\mathcal{D}_N\}$, and outputs two separated features of $F_{\mathcal{C}}^{\mathcal{D}}$ which is a domain-invariant feature and $F_{\mathcal{S}}^{\mathcal{D}}$, which are domain-related features as $(F_{\mathcal{C}}^{\mathcal{D}}, F_{\mathcal{S}}^{\mathcal{D}}) = E(I^{\mathcal{D}})$. By disentangling these two, we aim to effectively distinguish what to focus when conducting target task such as segmentation on an unseen domain.

For an image generator G, we take the StyleGAN2-based decoder [8] which is capable of generating high-quality images. Combination of Style-based generator with an encoder for conditional image generation or image translation is also showing a good performance on various works [15]. High-quality synthesized images with mixed domain information lead to improved disentanglement.

2.2 Loss Function

The generator aims to synthesize realistic images such as $\hat{I}^{(i,j)} = G(F_{\mathcal{C}}^{\mathcal{D}_i}, F_{\mathcal{S}}^{\mathcal{D}_j})$ that matches the distribution of the input style images, while maintaining the consistency of a content features. For this, reconstruction loss term is introduced first to maintain self-consistency:

$$\mathcal{L}_{rec} = |I^{\mathcal{D}_i} - \hat{I}^{(i,j)}|_1 \quad \text{for} \quad i = j \tag{2}$$

Also, adversarial loss used to train G to synthesize a realistic images.

$$\mathcal{L}_{adv} = -log(D(\hat{I}^{(i,j)})), \ \forall i, j \tag{3}$$

Finally, segmentor S tries to conduct a main task which should be work well on the unseen target domain. To enable this, segmentation decoder only focuses on the disentangled content feature rather than the style features, as in Fig. 1. To utilize high-resolution information, skip connections from an encoder is fed to a segmentation decoder. Segmentation decoder computes losses between segmentation predictions $\hat{y} = S(F_{\mathcal{C}}^{\mathcal{D}})$ and an segmentation annotation y. We use dice loss functions for segmentation task, as:

$$\mathcal{L}_{seg} = 1 - \frac{2|\hat{y} \cap y|}{|\hat{y}| + |y|} \tag{4}$$

To better separate domain-invariant contents from domain-specific features, a patch-wise adversarial discriminator PD is included in the training, in a similar manner as introduce in [15]. With an adversarial loss between the patches within an image and between translated images, the encoder is trained to better disentangle styles of a domain as below. The effectiveness of the loss is compared on the experiment session.

$$\mathcal{L}_{padv} = -log(PD(\hat{I}^{(i,j)})) \quad \text{for } i \neq j. \tag{5}$$

Under an assumption that well-trained disentangled representation learning satisfies the identity on content features, we apply an identity loss on both contents and segmentation outputs for a translated images. In addition to a regularization effect, this leads to increased performance and a stability in the training.

$$\mathcal{L}_{identity} = |F_{\mathcal{C}}^{\mathcal{D}_i}, F_{\mathcal{C}}^{\mathcal{D}_i *}|_1 + (1 - \frac{2|\hat{y}^* \cap y|}{|\hat{y}^*| + |y|}), \tag{6}$$

where $(F_{\mathcal{C}}^{\mathcal{D}_i *}, F_{\mathcal{S}}^{\mathcal{D}_j *}) = E(\hat{I}^{(i,j)})$, and $\hat{y}^* = S(F_{\mathcal{C}}^{\mathcal{D}_i *})$.

Therefore, overall loss function becomes as below.

$$\mathcal{L}_{all} = \lambda_0 \mathcal{L}_{seg} + \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{identity} + \lambda_3 \mathcal{L}_{adv} + \lambda_4 \mathcal{L}_{padv}, \tag{7}$$

where λ_0, λ_1, λ_2, λ_3, and λ_4 are the weights of $\mathcal{L}_{seg}$, $\mathcal{L}_{rec}$, $\mathcal{L}_{identity}$, $\mathcal{L}_{adv}$, and λ_{padv}, respectively.

3 Experiments and Results

To evaluate the effectiveness of our proposed approach in addressing domain generalization for medical fields, we conduct experiments on a public dataset. The method is trained on three source domains and evaluates its performance on the remaining target domain. We compare our results to those of existing methods, including Fed-DG [11], DoFE [18], RAM-DSIR [22], and DCAC [5].

3.1 Setup

Dataset. We evaluate our proposed method on a public dataset, called Fundus [18]. The **Fundus** dataset consists of retinal fundus images for optic cup (OC) and disc (OD) segmentation from four medical centers. Each of four domains has 50/51, 99/60, 320/80 and 320/80 samples for each training and test. For all results from the proposed method, the images are randomly cropped for these data, then resized the cropped region to 256 × 256. The images are normalized to a range of −1 to 1 using the min-max normalization and shift process.

Metrics. We adopt the Dice coefficient (Dice), a widely used metric, to assess the segmentation results. Dice is calculated for the entire object region. A higher Dice coefficient indicates better segmentation performance.

3.2 Implementation Details

Our proposed network is implemented based on 2D. The encoder network is applied with four and three downsampling layers and with channel numbers of 32, 64, 160, and 256. Adam optimizer with a learning rate of 0.0002 and momentum of 0.9 and 0.99 is used. We set λ_0, λ_1, λ_2, λ_3, and λ_4 as 0.1, 1, 0.5, 0.5, and 0.1, respectively. We train the proposed method for 100 epochs on the

Table 1. Quantitative results of Dice for Optic Cups (OC) and Optic Discs (OD) segmentation on Fundus segmentation task. The highest results are **bolded**.

Task	Optic Cup segmentation					Optic Disc Segmentation					Total
Unseen Site	A	B	C	D	Avg	A	B	C	D	Avg	**Avg**
Baseline	75.60	77.11	80.74	81.29	78.69	94.96	85.30	92.00	91.11	90.84	84.76
FedDG	82.96	72.92	84.09	83.21	80.80	95.00	87.44	91.89	92.06	91.60	86.21
DoFE	82.34	**81.28**	86.34	84.35	83.58	95.74	89.25	93.6	93.9	93.12	88.35
RAM-DSIR	84.28	80.17	**86.72**	84.12	83.24	94.81	88.05	**94.78**	93.13	92.66	88.26
DCAC	82.79	75.72	86.31	**86.12**	82.74	**96.26**	87.51	94.13	**95.07**	**93.24**	87.99
Baseline+L_{rec}+L_{adv}	76.97	73.10	83.80	83.36	79.31	94.96	87.89	93.26	93.25	92.34	85.82
Baseline+L_{rec}+L_{adv}+L_{padv}	80.94	73.55	84.57	85.12	81.05	95.46	87.38	93.02	91.80	91.92	86.48
Ours	**85.70**	79.00	86.43	85.40	**84.13**	96.00	**89.50**	93.45	92.80	92.94	**88.54**

Fundus dataset with a batch size of 6. Each batch consists of 2 slices from each of the three domains. We use data augmentation to expand the training samples, including random gamma correction, random rotation, and flip. The training is implemented on one NVIDIA RTX A6000 GPU.

3.3 Results

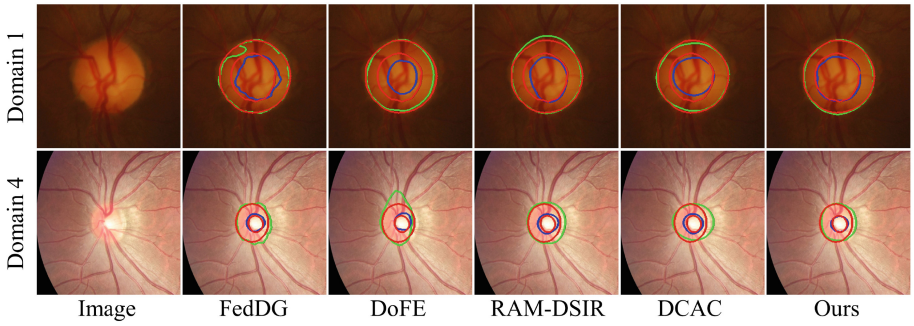

Fig. 2. Results figures for two domains. Red circles indicates ground truth, and blue and green each indicates predictions of OC and OD for each methods. (Color figure online)

We evaluate the performance of our proposed method by comparing it to four existing methods, as mentioned earlier. Table 1 shows quantitative results of Dice coefficient. Except for our method, DCAC has shown effectiveness in generalization with an average Dice score of 82.74 and 93.24 for OC and OD, respectively. Our proposed approach demonstrates impressive and effective results across all evaluation metrics. Our proposed method also performs effectively, with an average Dice score of 84.03 and 92.94 for OC and OD, respectively. It demonstrates that our method performs effectively compared to the previous methods. We

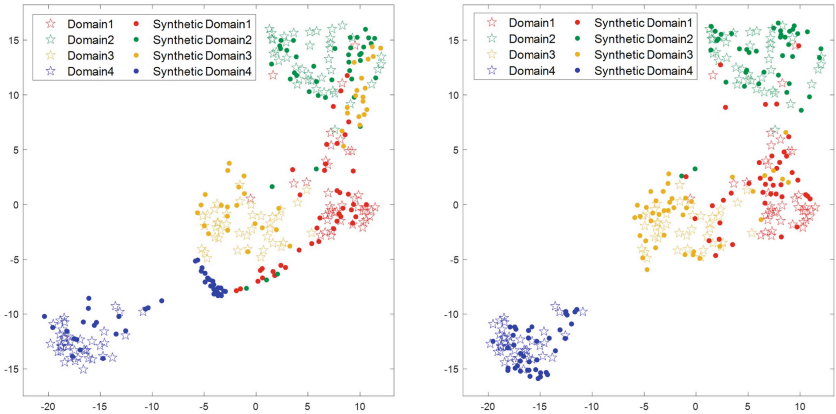

Fig. 3. t-SNE visualization with raw data. (a) is visualized about raw data and amplitude mixing derived in RAM-DSIR [22]. (b) is based on the raw data and the style mixing, which are mixed using our proposed method.

also perform qualitative comparisons with the other methods, as shown in Fig. 2. Specifically, for an image in Domain 1, as depicted in the first row, the boundary of OC is difficult to distinguish. The proposed method accurately identifies the exact regions of both OC and OD regions.

Furthermore, we analyze ablation studies to evaluate the effectiveness of each term in our loss function, including $\mathcal{L}_{seg}$, $\mathcal{L}_{rec}$, $\mathcal{L}_{identity}$, $\mathcal{L}_{adv}$, and $\mathcal{L}_{padv}$. The impact of adding each term sequentially on performance is analyzed, and the results are presented in Table 1. Our findings indicate that each term in the loss function plays a crucial role in generalizing domain features. As each loss term is added, we observed a gradual increase in the quantitative results for all unseen domains. For instance, when the unseen domain is Domain A, the Dice score improved from 75.60 to 76.97 to 85.70 upon adding each loss term.

To evaluate the effectiveness of our model in extracting variant and invariant features, we conducted t-SNE visualization on the style features of images synthesized using two different methods: the widely-used mixup method as in [22] and our proposed method. As illustrated in Fig. 3, the images generated by our model exhibit enhanced distinguishability compared to the mixup-based mixing visualization. This observation suggests that the distribution of mixed images using our method closely aligns with the original domain, which is better than the mixup-based method. It indicates that our model has successfully learned to extract both domain-variant and domain-invariant features through disentanglement learning, thereby contributing to improved generalizability.

4 Conclusion

Our work addresses the challenge of domain generalization in medical image segmentation by proposing a novel framework for retinal fundus image segmentation. The framework leverages disentanglement learning with adversarial and regularized training to extract invariant features, resulting in significant improvements over existing approaches. Our approach demonstrates the effectiveness of leveraging domain knowledge from multiple sources to enhance the generalization ability of deep neural networks, offering a promising direction for future research in this field.

Acknowledgements. This research was supported by Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Science and ICT (2021R1A4A1031437, 2022R1A2C2008983, 2021R1C1C2008773), Artificial Intelligence Graduate School Program at Yonsei University [No. 2020-0-01361], the KIST Institutional Program (Project No.2E32271-23-078), and partially supported by the Yonsei Signature Research Cluster Program of 2023 (2023-22-0008).

References

1. Akada, H., Bhat, S.F., Alhashim, I., Wonka, P.: Self-supervised learning of domain invariant features for depth estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3377–3387 (2022)
2. 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)
3. Hinz, T., Wermter, S.: Image generation and translation with disentangled representations. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2018)
4. Hoffman, J., et al.: Cycada: cycle-consistent adversarial domain adaptation. In: International Conference on Machine Learning, pp. 1989–1998. PMLR (2018)
5. Hu, S., Liao, Z., Zhang, J., Xia, Y.: Domain and content adaptive convolution based multi-source domain generalization for medical image segmentation. IEEE Trans. Med. Imaging **42**(1), 233–244 (2022)
6. Hu, Y., Ma, A.J.: Adversarial feature augmentation for cross-domain few-shot classification. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13680, pp. 20–37. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20044-1_2
7. 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)
8. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of styleGAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110–8119 (2020)
9. Lee, S., Cho, S., Im, S.: Dranet: disentangling representation and adaptation networks for unsupervised cross-domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15252–15261 (2021)

10. Li, H., Wang, Y., Wan, R., Wang, S., Li, T.Q., Kot, A.: Domain generalization for medical imaging classification with linear-dependency regularization. In: Advances in Neural Information Processing Systems, vol. 33, pp. 3118–3129 (2020)
11. Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P.A.: Feddg: federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1013–1023 (2021)
12. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
13. Lu, W., Wang, J., Li, H., Chen, Y., Xie, X.: Domain-invariant feature exploration for domain generalization. arXiv preprint arXiv:2207.12020 (2022)
14. Park, D., et al.: Importance of CT image normalization in radiomics analysis: prediction of 3-year recurrence-free survival in non-small cell lung cancer. Eur. Radiol. 1–10 (2022)
15. Park, T., et al.: Swapping autoencoder for deep image manipulation. In: Advances in Neural Information Processing Systems, vol. 33, pp. 7198–7211 (2020)
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. Shin, H., Kim, H., Kim, S., Jun, Y., Eo, T., Hwang, D.: SDC-UDA: volumetric unsupervised domain adaptation framework for slice-direction continuous cross-modality medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7412–7421 (2023)
18. Wang, S., Yu, L., Li, K., Yang, X., Fu, C.W., Heng, P.A.: Dofe: domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. IEEE Trans. Med. Imaging **39**(12), 4237–4248 (2020)
19. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems, vol. 34, pp. 12077–12090 (2021)
20. Zhang, B., et al.: StyleSwin: transformer-based GAN for high-resolution image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11304–11314 (2022)
21. Zhang, L., et al.: Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE Trans. Med. Imaging **39**(7), 2531–2540 (2020)
22. Zhou, Z., Qi, L., Shi, Y.: Generalizable medical image segmentation via random amplitude mixup and domain-specific image restoration. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13681, pp. 420–436. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19803-8_25
23. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)