IntraStyler: Intra-Domain Style Synthesis for Cross-Modality MRI Domain Adaptation

Han Liu[1,2]✉, Yubo Fan[2], Hao Li[2], Dewei Hu[3], Daniel Moyer[2], Zhoubing Xu[4], Benoit Dawant[2], and Ipek Oguz[2]

[1] Siemens Healthineers, Princeton, NJ, USA
[2] Vanderbilt University, Nashville, TN, USA
[3] Mayo Clinic, Rochester, MN, USA
[4] Johnson & Johnson Innovative Medicine, Titusville, NJ, USA
han.liu@siemens-healthineers.com

Abstract. Segmentation of vestibular schwannoma and cochlea from T2 MRI is clinically important yet annotation-intensive. Domain adaptation (DA) has been widely adopted to bridge the gap between labeled contrast-enhanced T1 and unlabeled T2 datasets. While existing methods focus on cross-domain alignment, intra-domain variability within the target domain remains largely overlooked. Images from the same domain may vary substantially due to different scanners, field strengths, and acquisition protocols. Ignoring this variability produces homogeneous synthetic images that limit the generalizability of downstream segmentation models. To address this, we propose ***IntraStyler***, a 3D unpaired image translation method that automatically discovers fine-grained intra-domain styles without any predefined sub-domains, and synthesizes diverse target domain images using per-image style references. To this end, we design a 3D style encoder trained with a novel contrastive learning objective to extract style-only embeddings disentangled from anatomy. IntraStyler is built upon the 1st place CrossMoDA challenge solution and further advances it, generating more diverse synthetic data and achieving more reliable downstream segmentation. Code is available at https://github.com/MedICL-VU/IntraStyler.

Keywords: domain adaptation, style transfer, contrastive learning

1 Introduction

Vestibular schwannoma (VS) is a benign tumor arising from the vestibular nerve, and accurate segmentation of VS and cochlea from MRI is essential for treatment planning and follow-up. While contrast-enhanced T1 (ceT1) MRI is the standard modality, high-resolution T2 imaging offers a lower-risk and more accessible alternative [2]. However, annotating T2 images is costly and time-consuming, motivating domain adaptation (DA) approaches that transfer knowledge from labeled ceT1 to unlabeled T2 without target domain annotations [4,20]. One of the most effective DA strategies is image-level domain alignment [3,16,19], where source

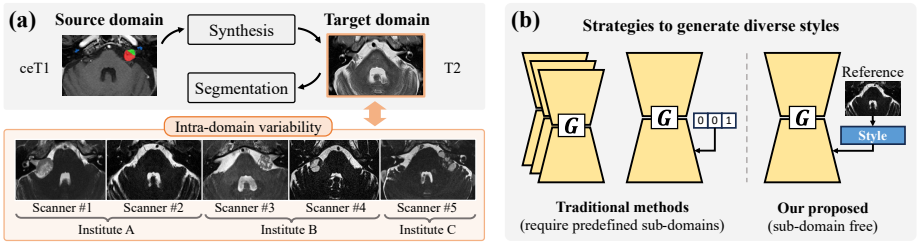

Fig. 1. (a) Intra-domain variability: images from the same domain exhibit diverse appearances across institutes and scanners. **(b)** Comparison of style generation strategies.

domain images are synthesized into the target domain via unpaired image translation [21,17], and the synthetic images are used with source domain labels to train the segmentation model. Classical DA assumes images within the same domain share the same distribution. However, in many real-world medical datasets, images from the same domain (e.g., T2 MRI) may span multiple sub-domains arising from differences in scanners or acquisition protocols, also known as ***intra-domain*** variability. If DA methods fail to capture this variability and produce only homogeneous synthetic images, the downstream segmentation model will have poor generalizability across the target domain.

Most existing methods require **predefined sub-domains** to synthesize diverse styles. This assumption has two limitations in practice. First, sub-domain labels are not always available. Second, coarse sub-domain definitions fail to capture fine-grained appearance variation. As shown in Fig. 1, previous studies use *institutes* as sub-domains, yet images from the same institute may still differ substantially due to varying *scanners* or *protocols*. [15,22] train separate synthesis networks per source and target sub-domain pair, which is highly inefficient and degrades when sub-domain data is scarce. [16,7,1,14] train a unified network conditioned on predefined sub-domains, which still fundamentally depends on their availability and accuracy. It is therefore desirable to develop a method that captures intra-domain variability without any predefined sub-domains.

In this paper, we present ***IntraStyler***, an unpaired image translation method that captures diverse intra-domain styles without requiring any sub-domain labels. Our key insight is that each target domain image implicitly encodes its own fine-grained style and can therefore serve as a style reference at inference time. To extract style information disentangled from anatomy, we design a 3D style encoder trained with a novel contrastive learning objective. Built upon the 1st place CrossMoDA solution [16], IntraStyler demonstrates that (1) sub-domain-free style discovery yields more diverse synthetic images and (2) improves downstream segmentation generalizability. Our contributions are as follows:

- We propose **IntraStyler**, the first 3D unpaired image translation method for cross-modality MRI domain adaptation that synthesizes diverse target domain styles without any predefined sub-domain labels.

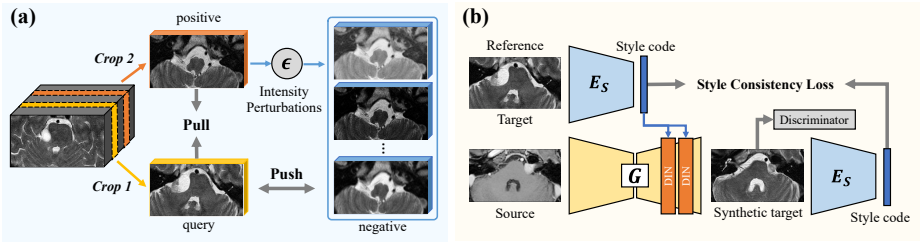

Fig. 2. **(a)** Contrastive learning setup for style encoder E_S: given a query, the positive shares its style but differs in anatomy, while negatives share the positive's anatomy but have different styles. **(b)** IntraStyler: the style encoder E_S is jointly trained with synthesis backbone G, where the style embedding extracted from a reference image is used to condition G to synthesize the corresponding style.

- We design a contrastive learning objective to extract per-image style embeddings from 3D MRI that are disentangled from anatomical content, enabling fine-grained style conditioning and smooth style interpolation.
- Evaluated on the CrossMoDA challenge benchmark, IntraStyler generates more diverse synthetic styles and achieves superior downstream segmentation performance compared to the state-of-the-art method.

2 Methods

As shown in Fig. 2, IntraStyler consists of two components: (1) a style encoder trained by contrastive learning to extract style-only features from target domain images (Sec. 2.1), and (2) a synthesis network conditioned on the extracted style embedding for controllable image translation (Sec. 2.2).

2.1 Style Extraction via Contrastive Learning

We aim to train a 3D style encoder E_S that is sensitive to style changes but invariant to anatomical content. Our key intuition is that by training on artificially simulated style variations, the encoder learns a general notion of style that transfers to discriminating real MRI appearances without requiring any style annotations. As shown in Fig. 2(a), we define *query* as a 3D patch randomly cropped from a target domain image, *positive* as another patch from the same image at a different location, and *negatives* as intensity-perturbed versions of the positive. Thus, positive and negatives share the same anatomy but differ in style, while only the positive shares the style of the query.

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the source and target domains, and let y and y^+ denote the query and positive. The perturbation function $\epsilon(\cdot)$ is randomly sampled from a set of intensity transformations (e.g., contrast adjustment, Gaussian smoothing), giving negative $y^- = \epsilon(y^+)$. The query, positive, and N negatives are passed

through E_S to obtain K-dimensional style vectors v, $v^+ \in \mathbb{R}^K$ and $v^- \in \mathbb{R}^{N\times K}$, which are ℓ_2-normalized onto a unit sphere so that similarity reduces to the dot product. The contrastive objective is cast as an $(N+1)$-way classification problem, with cross-entropy loss maximizing the probability of selecting the positive over all negatives:

$$L_{\text{style}}(v, v^+, v^-) = -log[\frac{exp(v \cdot v^+/\tau)}{exp(v \cdot v^+/\tau) + \sum_{n=1}^{N} exp(v \cdot v_n^-/\tau)}] \tag{1}$$

where $\tau = 0.01$ is the temperature scaling factor for style similarity.

2.2 Controllable Synthesis with Extracted Styles

To integrate the extracted style with the synthesis model, we inject the style embedding as a condition for controllable synthesis, as shown in Fig. 2(b). Following [16], we adopt 3D QS-Attn [8] as the synthesis backbone, selected for its attention mechanism and one-sided design that is computationally efficient for 3D tasks. Inspired by adaptive normalization methods [5,13,12,9], the style embedding is injected via dynamic instance normalization (DIN) layers. The input feature z is first channel-wise normalized as in standard IN: $z_{\text{norm}} = \frac{z-\mu}{\sigma}$. The style vector v is then passed through a $1\times1\times1$ convolutional mapping layer to produce affine parameters γ_v and β_v, which de-normalize the feature as $z_{\text{out}} = \gamma_v z_{\text{norm}} + \beta_v$. In practice, the last two IN layers of the decoder are replaced with DIN layers.

To further encourage the consistency between the styles of reference image and the generated image, we use E_s to extract the style embedding of the generated image $G(x)$ and introduce a style consistency loss:

$$L_{\text{con}} = -sim(E_S(y), E_S(G(x))) \tag{2}$$

where $x \in \mathcal{X}$ is the source domain input image, G is the synthesis network, and $sim(\cdot)$ is the dot product, since the style embeddings are unit vectors.

Training Objective. Instead of training style encoder and synthesis network sequentially, we find these two networks can be jointly optimized in an end-to-end manner. The overall training objective of IntraStyler is expressed as:

$$L_{\text{IntraStyler}} = \underbrace{L_{\text{adv}} + \lambda_{\text{NCE}} L_{\text{PatchNCE}}}_{\text{QS-Attn}^5} + \underbrace{\lambda_{\text{style}} L_{\text{style}}}_{\text{style discrimination}} + \underbrace{\lambda_{\text{con}} L_{\text{con}}}_{\text{style consistency}} \tag{3}$$

[5] L_{adv} and L_{PatchNCE} are losses from CUT [17], inherited by QS-Attn [8]. See https://github.com/sapphire497/query-selected-attention.

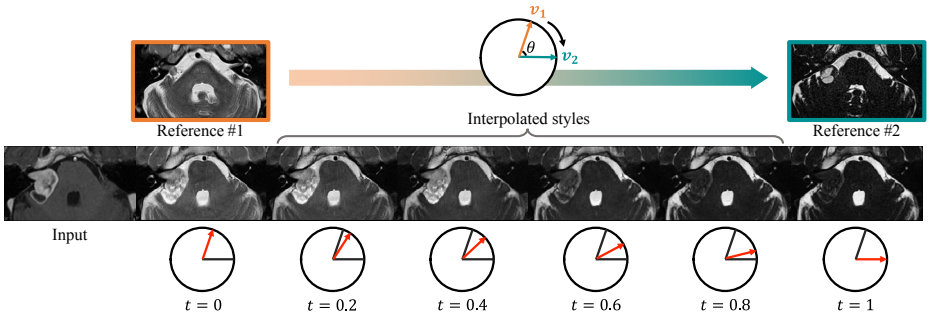

Fig. 3. As the interpolation parameter t varies from 0 to 1 (red arrow rotates clockwise), the *appearance* of the synthesized image smoothly transitions between the styles of two reference images, while the anatomical content remains determined by the source input.

2.3 Style Interpolation with SLERP

Beyond a single reference image, IntraStyler supports style synthesis conditioned on two reference images via style interpolation. Standard linear interpolation between unit vectors produces intermediate vectors of varying magnitude, causing inconsistent style intensities. We therefore adopt spherical linear interpolation (SLERP) [11], which preserves unit norm throughout interpolation and ensures perceptually uniform style transitions. As shown in Fig. 3, the interpolation parameter $t \in [0, 1]$ provides explicit control over the degree of style blending between the two references. Given unit-norm style embeddings v_0 and v_1 with angle $\theta = \arccos(v_0 \cdot v_1)$ between them, the interpolated style is computed as:

$$\mathrm{SLERP}(v_0, v_1; t) = \frac{\sin((1-t)\theta)}{\sin(\theta)} v_0 + \frac{\sin(t\theta)}{\sin(\theta)} v_1 \tag{4}$$

2.4 Implementation Details

We implement IntraStyler based on the 1st place CrossMoDA solution [16] by incorporating our proposed style encoder. We set $\lambda_{\text{style}} = \lambda_{\text{con}} = 5$ and the remaining loss weights follow the baseline implementation [16]. The style vector dimension is set to $K = 256$. For contrastive learning, negative samples are constructed by applying one of the following intensity transformations to the positive patch: random contrast adjustment, Gaussian smoothing, Gaussian noise, bias field corruption, or a random mixture of all four. These transformations simulate the dominant physical sources of MRI appearance variation (e.g., B1 inhomogeneity, sequence parameters, thermal noise), enabling the encoder to learn a broad notion of style that generalizes to real acquisition differences (as shown in Fig. 4). For each training iteration, $N = 8$ negative samples are generated by independently sampling a random transformation for each.

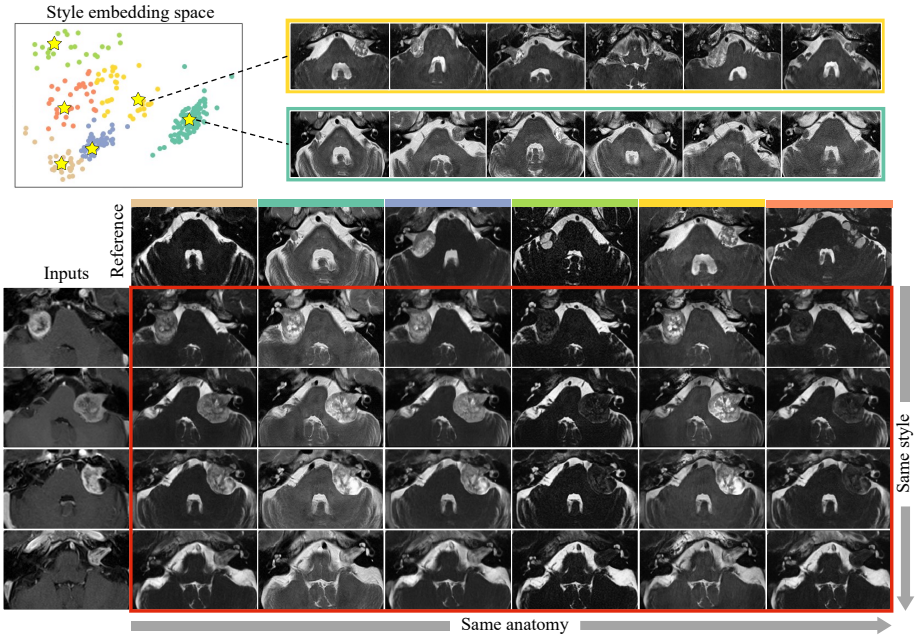

Fig. 4. Top left: the style embedding space of all target domain images. Clustering is done by K-means. **Top right**: visualization of samples from two clusters. Images within the same cluster share consistent styles despite varying anatomies, while images across clusters exhibit clearly distinct appearances. **Bottom**: source domain images (left column) synthesized to target domain using different reference images (colors of the references correspond to the cluster colors). Synthesis results are shown in the red.

3 Experiments and Results

We evaluate IntraStyler on the CrossMoDA challenge benchmark [20], which consists of 226 labeled ceT1 MRIs (source domain) and 295 unlabeled T2 MRIs (target domain), with segmentation masks for cochlea and intra- and extra-meatal components of vestibular schwannoma (VS). Data were collected from multiple institutes across scanners from four manufacturers (Siemens, Philips, GE, Hitachi), field strengths of 1.0T, 1.5T, and 3.0T, and multiple acquisition sequences, resulting in substantial intra-domain variability. This dataset is organized into only 3 coarse institute-based subsets, where the third subset itself aggregates data from multiple centers. Within each subset, images still vary substantially across scanner types and acquisition sequences, making institute labels an unreliable proxy for appearance variation. The test set consists of 96 T2 MRIs drawn from this heterogeneous multi-institute, multi-scanner cohort. Next, we investigate three key questions to validate the core components of IntraStyler:

Q1. Does contrastive learning extract style-only information from real-world heterogeneous MRIs? To assess our style embeddings, we use

Table 1. Quantitative results on segmentation performance. *Subdomain-free* indicates whether the method requires predefined sub-domain labels. ASSD follows the challenge evaluation protocol [20], where a failure is defined as the target structure being completely missed, and 350 mm is assigned to all failure cases. #Fail reports the number of failures per method. Shaded cells indicate a statistically significant difference from IntraStyler (Wilcoxon signed-rank, $p<0.05$). **Bold** = best per row.

		NoVar	MultiNets	Dynamic [16]	IntraStyler
Diverse styles		✗	✓	✓	✓
Unified model		✓	✗	✓	✓
Subdomain-free		✓	✗	✗	✓
Dice↑	Intra-VS	0.08±0.14	0.63±0.24	0.57±0.26	**0.69±0.12**
	Extra-VS	0.10±0.22	0.70±0.29	0.61±0.34	**0.81±0.12**
	Cochlea	0.81±0.03	0.82±0.07	0.82±0.04	**0.83±0.03**
ASSD↓	Intra-VS	219.38±169.51	22.50±85.01	30.46±96.89	**4.35±35.67**
	Extra-VS	222.41±168.24	28.92±95.53	53.11±125.16	**0.60±0.31**
	Cochlea	0.24±0.14	0.40±1.75	0.23±0.14	**0.21±0.12**
	Boundary	269.88±147.50	36.88±106.98	64.96±136.10	**0.65±0.33**
#Fail↓	Intra-VS	60	6	8	**1**
	Extra-VS	55	7	13	**0**
	Cochlea	0	0	0	0

the style encoder to project all target domain images into the style embeddings. Then we use K-means to find the clusters of similar style embeddings and visualize them with PCA. The number of clusters is determined by silhouette analysis. As shown in Fig. 4 (top), images within the same cluster share consistent styles despite varying anatomies, while images across clusters exhibit clearly distinct appearances, confirming that the learned embeddings capture style independently of anatomy. Notably, although the style encoder is trained only on artificially perturbed styles, it generalizes to discriminate real MRI styles.

Q2. Do synthesized images follow the style of the reference image? For style synthesis evaluation (Fig. 4, bottom), four source domain images are translated using six reference images, which are selected as the most representative sample of each cluster by [6]. We can observe that (1) column-wise, synthesized images are consistent with their reference style. (2) row-wise, the same source image produces visually distinct outputs across references while preserving anatomy. These results demonstrate that IntraStyler enables fine-grained style control without predefined sub-domains.

Q3. Does style diversity benefit downstream segmentation? We investigate the impact of different synthesis methods on the downstream segmentation task. For fair comparison, we compare four synthesis strategies that differ only in how intra-domain variability is handled, while keeping the downstream segmentation model fixed (nnU-Net [10]). We are aware of the existence of other DA approaches that do not require explicit image synthesis, e.g., feature-level alignment. *Comparison against these methods is beyond the scope of our study*

because (1) our major contribution lies specifically in style diversification during image translation and (2) IntraStyler is implemented based on the challenge-winning method [16], which already outperformed other DA strategies.

We compare the following synthesis strategies: (1) **Dynamic** [16]: the 1st place solution of the CrossMoDA challenge, which generates diverse styles by conditioning a unified network on 3 predefined institute-based sub-domains, represented as one-hot codes (e.g., $[1, 0, 0]$). (2) **MultiNets**: shares the same backbone as Dynamic but replaces the unified design with separate synthesis networks trained for each of the 3 institute-based sub-domains, and thus learns an easier task per network at the cost of reduced efficiency. (3) **NoVar**: uses the same backbone as Dynamic but without any intra-domain modeling, serving as a lower bound to quantify the impact of ignoring target domain variability. (4) **IntraStyler**: to avoid selecting reference images with repetitive styles, we project all target domain images into the learned style embedding space and apply K-means clustering with silhouette analysis to identify meaningful style groups. Within each cluster, the most representative image is selected using the diversity-based method of [6]. This yields 7 reference images that capture the majority of style diversity in the target domain. Compared to sub-domain-based methods that produce only 3 distinct styles, IntraStyler generates a richer set of synthetic target domain images for downstream segmentation training.

We evaluate segmentation results using Dice and Average Symmetric Surface Distance (ASSD) across Intra-VS, Extra-VS, and Cochlea, with Boundary ASSD (i.e., the distance between the intra- and extra-VS boundary) additionally reported. Following the challenge evaluation protocol [20], a failure is defined as the target structure being completely missed, and a 350 mm penalty is assigned to all failure cases in the ASSD computation. We report the quantitative results in Tab. 1 with statistical significance between each baseline and IntraStyler using the Wilcoxon signed-rank test with Bonferroni correction ($p<0.05$).

NoVar performs significantly worse than IntraStyler across all structures, with 60 and 55 complete failures on Intra-VS and Extra-VS respectively, confirming that data augmentation alone is insufficient to address real-world intra-domain variability and model-based style synthesis is necessary. Notably, all methods achieve zero Cochlea failures, suggesting that intra-domain variability primarily affects VS segmentation. Among sub-domain-based methods, MultiNets is the strongest competitor yet still produces substantially more failures (6 vs. 1 on Intra-VS, 7 vs. 0 on Extra-VS) with significantly lower Dice on Extra-VS and Cochlea. Dynamic, the prior challenge winning method, similarly underperforms IntraStyler significantly across nearly all Dice and ASSD metrics. Both methods share a fundamental limitation: institute-level sub-domains are too coarse to capture true intra-domain variation, as images from the same institute can differ substantially across scanners and protocols. IntraStyler overcomes this by automatically discovering fine-grained style clusters within the target domain, achieving the best performance across all reported metrics.

4 Discussion and Conclusion

IntraStyler demonstrates that predefined sub-domain labels are not necessary for diverse style synthesis in cross-modality DA. By treating each target domain image as an independent style reference, the method naturally scales to any degree of intra-domain heterogeneity without additional prior information. IntraStyler has two limitations. First, the intensity perturbation functions for negative sample construction are designed for MR and and may require modality-specific tuning for other imaging modalities, such as CT and PET. Second, the global style vector from average pooling discards spatial information, preventing control of local style variations such as tumor-specific textures. Our global style conditioning could be combined with local tumor augmentation approaches [18] to further improve robustness. This work opens several future research directions. First, the learned style embedding space exhibits meaningful clustering without any supervision, suggesting its potential utility for unsupervised scanner harmonization and image retrieval. Second, while IntraStyler is currently built on a GAN-based backbone, the style conditioning mechanism is architecture-agnostic and its integration with diffusion models or flow matching models is a promising direction for higher-fidelity synthesis. We believe IntraStyler provides a practical foundation for intra-domain variability in real-world DA settings where sub-domain annotations are often unavailable.

References

1. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8188–8197 (2020)
2. Coelho, D.H., Tang, Y., Suddarth, B., Mamdani, M.: Mri surveillance of vestibular schwannomas without contrast enhancement: clinical and economic evaluation. The Laryngoscope **128**(1), 202–209 (2018)
3. Dong, H., Yu, F., Zhao, J., Dong, B., Zhang, L.: Unsupervised domain adaptation in semantic segmentation based on pixel alignment and self-training. arXiv preprint arXiv:2109.14219 (2021)
4. Dorent, R., Kujawa, A., Ivory, M., Bakas, S., Rieke, N., Joutard, S., Glocker, B., Cardoso, J., Modat, M., Batmanghelich, K., et al.: Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. Medical Image Analysis **83**, 102628 (2023)
5. Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style. In: International Conference on Learning Representations (2017), https://openreview.net/forum?id=BJO-BuT1g
6. Hacohen, G., Dekel, A., Weinshall, D.: Active learning on a budget: Opposite strategies suit high and low budgets. arXiv preprint arXiv:2202.02794 (2022)
7. Han, L., Tan, T., Mann, R.: Fine-grained unsupervised cross-modality domain adaptation for vestibular schwannoma segmentation. In: International Challenge on Cross-Modality Domain Adaptation for Medical Image Segmentation, pp. 364–371. Springer (2023)

8. Hu, X., Zhou, X., Huang, Q., Shi, Z., Sun, L., Li, Q.: Qs-attn: Query-selected attention for contrastive learning in i2i translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18291–18300 (2022)
9. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
11. Jafari, M., Molaei, H.: Spherical linear interpolation and bézier curves. General Scientific Researches **2**(1), 13–17 (2014)
12. Jing, Y., Liu, X., Ding, Y., Wang, X., Ding, E., Song, M., Wen, S.: Dynamic instance normalization for arbitrary style transfer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 4369–4376 (2020)
13. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
14. Li, H., Liu, H., von Busch, H., Grimm, R., Huisman, H., Tong, A., Winkel, D., Penzkofer, T., Shabunin, I., Choi, M.H., et al.: Deep learning–based unsupervised domain adaptation via a unified model for prostate lesion detection using multisite biparametric mri datasets. Radiology: Artificial Intelligence **6**(5), e230521 (2024)
15. Liu, H., Fan, Y., Oguz, I., Dawant, B.M.: Enhancing data diversity for self-training based unsupervised cross-modality vestibular schwannoma and cochlea segmentation. In: International MICCAI Brainlesion Workshop. pp. 109–118. Springer (2022)
16. Liu, H., Fan, Y., Xu, Z., Dawant, B.M., Oguz, I.: Learning site-specific styles for multi-institutional unsupervised cross-modality domain adaptation. In: International Challenge on Cross-Modality Domain Adaptation for Medical Image Segmentation, pp. 372–385. Springer (2023)
17. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 319–345. Springer (2020)
18. Sallé, G., Conze, P.H., Bert, J., Boussion, N., Visvikis, D., Jaouen, V.: Cross-modal tumor segmentation using generative blending augmentation and self training. IEEE Transactions on Biomedical Engineering **71**(8), 2428–2439 (2024)
19. Shin, H., Kim, H., Kim, S., Jun, Y., Eo, T., Hwang, D.: Cosmos: Cross-modality unsupervised domain adaptation for 3d medical image segmentation based on target-aware domain translation and iterative self-training. arXiv preprint arXiv:2203.16557 (2022)
20. Wijethilake, N., Dorent, R., Ivory, M., Kujawa, A., Cornelissen, S., Langenhuizen, P., Okasha, M., Oviedova, A., Dong, H., Kang, B., et al.: crossmoda challenge: Evolution of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation from 2021 to 2023. arXiv preprint arXiv:2506.12006 (2025)
21. 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)
22. Zhuang, Y.: A 3d multi-style cross-modality segmentation framework for segmenting vestibular schwannoma and cochlea. arXiv preprint arXiv:2311.11578 (2023)