MSDG-StyleNet: Multi-source Unsupervised Domain-Generalized CBCT-to-CT Translation with Style-Consistent Disentangled Representations

Xin Long[1], Xinrui Liu[2], and Fan Gan[3]

[1] School of Mathematics and Computer Science, Nanchang University, Nanchang, China
[2] Queen Mary College of Nanchang University, Nanchang, China
[3] The Affiliated Eye Hospital, Jiangxi Medical College, Nanchang University, Nanchang, China

Abstract. Cone-beam computed tomography (CBCT) is gaining prominence in clinical radiology, particularly for intraoperative guidance, owing to its lower radiation dose and faster acquisition speed compared to computed tomography (CT). However, CBCT images often exhibit compromised quality, characterized by increased noise, artifacts, and diminished soft-tissue contrast, which can hinder their direct clinical application. While CBCT-to-CT translation presents a promising solution, this task faces significant challenges in multi-institutional settings where diverse imaging protocols introduce substantial domain shifts, especially when paired CBCT-CT data is scarce. Current unsupervised domain generalization (UDG) techniques often struggle to simultaneously maintain robust anatomical accuracy and preserve domain-specific characteristics—both crucial for clinical reliability. To address these limitations, we propose a novel disentangled representation learning framework for UDG-based CBCT-to-CT translation. Our method uniquely separates domain-invariant anatomical content from domain-specific styles, while leveraging learnable domain-style prototypes to dynamically capture key stylistic characteristics. To ensure high-quality translation, we implement a dual-level consistency mechanism that guarantees both anatomical fidelity and style alignment. By utilizing unpaired data for training and enabling flexible content-prototype combinations, our framework effectively generalizes to new institutions without requiring paired data. Extensive validation across three distinct institutional domains demonstrates that our method achieves superior anatomical accuracy and style fidelity compared to state-of-the-art approaches, establishing a clinically practical UDG paradigm with inherent cross-institutional interoperability.

Keywords: CBCT-to-CT Translation · Unsupervised Domain Generalization · Decoupled Representation Learning · Style Consistency.

1 Introduction

Cone-beam Computed Tomography (CBCT) has gained widespread adoption in clinical practice due to its lower radiation dose, reduced cost, and real-time imaging capabilities compared to conventional CT [1, 2]. However, CBCT images often suffer from artifacts, noise, and intensity inconsistencies across different scanning protocols and institutions, limiting their direct use in diagnosis and treatment planning [3, 4]. Converting CBCT to CT-like images while maintaining anatomical accuracy and achieving consistent quality across different centers remains a significant challenge in medical imaging.

Recent advances in image-to-image translation [5, 6, 7], particularly with deep learning approaches such as U-Net architectures [8] and Generative Adversarial Networks (GANs) [9], while providing baseline solutions [10], [11], lack the flexibility and generalization capabilities required for robust multi-center applications [12].

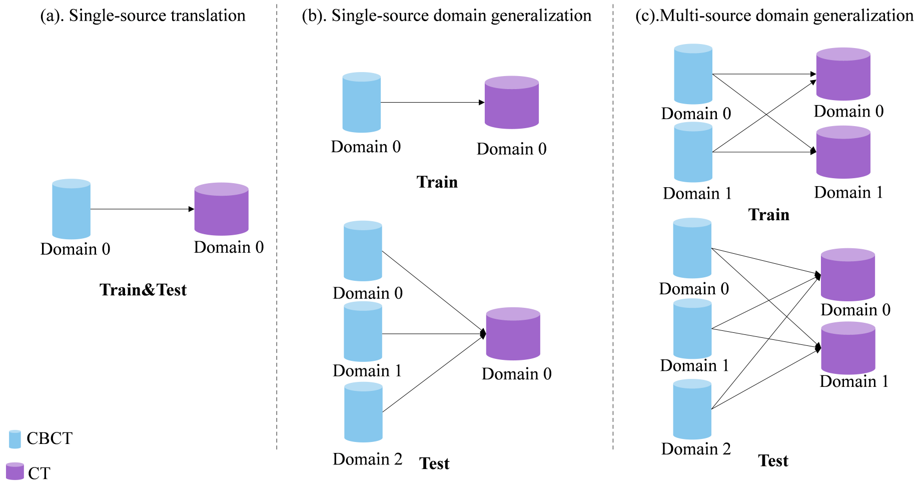

Fig. 1. Diagram of different approaches to CBCT-to-CT translation. (a) Single-source translation involves training and testing on data from the same institution. (b) Single-source domain generalization allows for training on one domain while testing on the same domain. (c) Multi-source domain generalization incorporates data from multiple centers, enhancing model robustness across diverse imaging conditions. Blue and purple colors represent CBCT and CT images, respectively.

To address domain shifts caused by protocol and scanner variations, unsupervised domain generalization (UDG) techniques have emerged [13, 14], employing strategies such as data augmentation [15, 16], domain-invariant feature learning [17] and meta-learning [18]. While these approaches enhance model robustness to domain variations [19], they often compromise on the precise anatomical detail preservation and high-fidelity reconstruction necessary for clinical CBCT-to-CT translation tasks.

Style transfer and disentangled representation learning [20] have been explored for medical image harmonization and modality synthesis [21]. However, existing methods typically employ simplified style guidance mechanisms and lack explicit domain generalization components, resulting in suboptimal performance when handling institution-specific imaging characteristics and cross-center variations in CBCT-to-CT translation.

To address these limitations, we propose MSDG-StyleNet, a novel framework that combines domain-style prototypes with disentangled representation learning for robust cross-institutional CBCT-to-CT translation. Our main contributions include:

1. We develop a novel framework enabling unpaired, unsupervised CBCT-to-CT translation with cross-institutional generalization capability.
2. We introduce a domain-style prototype learning mechanism that facilitates targeted style transfer to specific institutional domains during inference.
3. We establish a dual-level consistency mechanism by integrating disentangled representation learning with cycle consistency, ensuring robust content-style alignment.

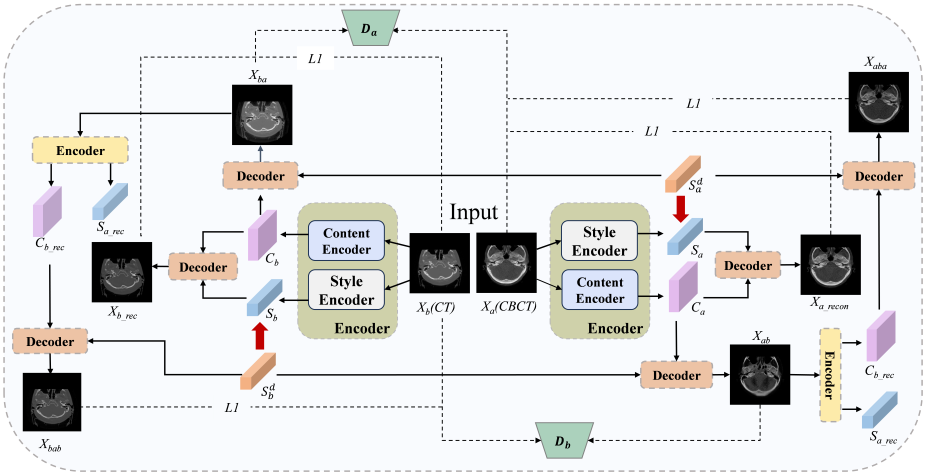

Fig. 2. Overview of the proposed MSDG-StyleNet architecture for unpaired CBCT-to-CT image translation. The network employs modality-specific content and style encoders to disentangle anatomical features and domain characteristics, respectively. Learnable domain-style prototypes capture domain-specific style information across different medical centers, while the shared decoder with AdaIN modules enables both self-reconstruction and cross-domain translation. Red arrows in the figure indicate the flow of parameter gradient updates during network optimization.

2 Method

2.1 Overview of the Architecture

We introduce the Multi-Source Domain Generalization with Style-Consistent Disentangled Network (MSDG-StyleNet) to address unpaired CBCT-to-CT image translation, specifically tackling domain shift between modalities X_a and X_b from different medical centers. Given $X_a = \{X_{a_1}, X_{a_2}, \ldots, X_{a_n}\}$ and $X_b = \{X_{b_1}, X_{b_2}, \ldots, X_{b_m}\}$, MSDG-StyleNet learns a mapping $f : X_a \rightarrow X_b$ without paired data.

As illustrated in Fig. 2, MSDG-StyleNet's architecture is composed of key components for disentangling content and style. Specifically, modality-specific content encoders, $E^c_{CBCT} : X_a \rightarrow C$ and $E^c_{CT} : X_b \rightarrow C$, map CBCT (X_a) and CT (X_b) images to a shared latent content space C, capturing modality-invariant semantic information. Similarly, modality-specific style encoders, $E^s_{CBCT} : X_a \rightarrow S$ and $E^s_{CT} : X_b \rightarrow S$, extract modality-specific style features by mapping images to a shared latent style space S, representing modality-specific textures and appearances. A shared decoder network $G : (C, S) \rightarrow X$ reconstructs images by integrating content codes from C and style codes from S, utilizing Adaptive Instance Normalization (AdaIN) modules for effective fusion, as detailed in Fig. 3(b). Furthermore, MSDG-StyleNet incorporates learnable domain-style prototypes, $S^d_a \in S$ and $S^d_b \in S$ for CBCT and CT respectively, serving as reference style prototypes for each modality.

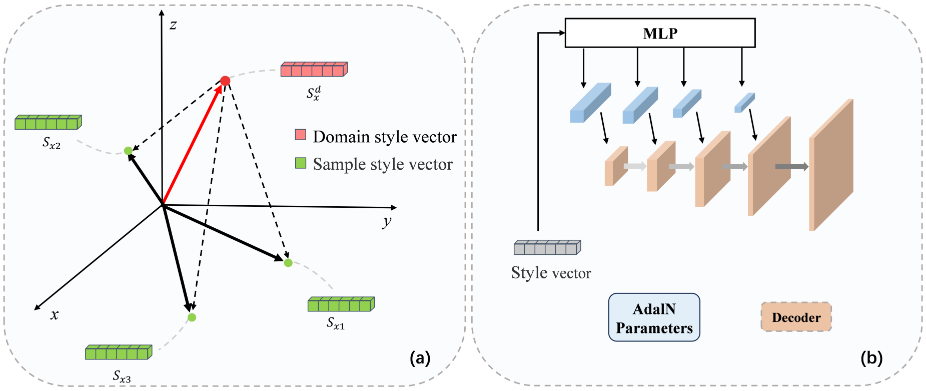

Fig. 3. (a) Illustration of our learnable style prototypes in the shared latent style space, where the red point (S^d_x) denotes the domain-style vector representing the overall stylistic essence of a domain, and green points (S_{x1}, S_{x2}, S_{x3}) depict sample-specific style vectors. These vectors are learned to effectively capture both global domain characteristics and individual variations. (b) The extracted style vector feeds into an MLP to generate AdaIN parameters, which are then used by the decoder to fuse content and style for image reconstruction or cross-domain translation.

2.2 Learnable Domain-Style Prototypes

For robust and generalized style transfer, domain-style prototypes are optimized through a carefully designed process. Initialized randomly in the shared latent style space S, as visualized in Fig. 3(a), these prototypes are iteratively refined using a domain loss function. While image-specific style codes (S_a, S_b) capture individual image characteristics, the prototypes S_a^d and S_b^d learn to represent fundamental stylistic patterns inherent to each modality. This optimization enables the prototypes to generalize style representation across varying scanning conditions and parameters within each domain.

The effectiveness of these prototypes is maintained through a dual mechanism that facilitates cross-domain translation and enables secondary reconstruction for accurate domain style representation. The validity of the prototypes is further reinforced through a dual-consistency constraint system. Cycle consistency loss preserves content integrity, while domain loss ensures style alignment. Additionally, adversarial learning guides prototype updates by supervising the domain adherence of the generated images.

2.3 Optimization Objective

We train MSDG-StyleNet using a composite loss function consisting of four components to ensure realistic image generation, content preservation, and effective style transfer between domains.

To make the generated cross-stylized images perceptually indistinguishable from real images, we employ an adversarial loss with two discriminators, D_a for CBCT and D_b for CT:

$$L^a_{adv} = \mathbb{E}_{x \in X_a}[\log D_a(x)] + \mathbb{E}_{y \in X_b}[\log(1 - D_a(G(C_b, S^d_a)))], \tag{1}$$

$$L^b_{adv} = \mathbb{E}_{y \in X_b}[\log D_b(y)] + \mathbb{E}_{x \in X_a}[\log(1 - D_b(G(C_a, S^d_b)))], \tag{2}$$

$$L_{adv} = L^a_{adv} + L^b_{adv}, \tag{3}$$

where $G(\cdot, \cdot)$ generates cross-stylized images and D_a, D_b measure real-image classification probabilities.

For content preservation, we incorporate a reconstruction loss using L1 norm to ensure reconstructed images maintain their original features:

$$\begin{aligned} L_{rec} = \mathbb{E}_{x,y}[&||x - G(E^c_{CBCT}(x), E^s_{CBCT}(x))||_1 \\ &+ ||y - G(E^c_{CT}(y), E^s_{CT}(y))||_1]. \end{aligned} \tag{4}$$

To maintain content integrity across style transfer cycles, we implement a cycle consistency loss that compares original and re-encoded content codes:

$$\begin{aligned} L_{cyc} = \mathbb{E}_{x,y}[&||x - G(E^c_{CBCT}(x), S^d_b)||_1 \\ &+ ||y - G(E^c_{CT}(y), S^d_a)||_1]. \end{aligned} \tag{5}$$

Finally, we ensure style consistency between image-specific codes and domain prototypes through a domain loss combining MSE and cosine similarity:

$$\begin{aligned} L_{dom} &= MSE(S_a, S_a^d) + MSE(S_b, S_b^d) \\ &\quad + \lambda_{cos}(2 - cos(S_a, S_a^d) - cos(S_b, S_b^d)). \end{aligned} \tag{6}$$

These components are combined into a total loss function with balanced weights:

$$L_{total} = \lambda_{adv} L_{adv} + \lambda_{rec} L_{rec} + \lambda_{cyc} L_{cyc} + \lambda_{dom} L_{dom}, \tag{7}$$

where λ_{adv}, λ_{rec}, λ_{cyc}, and λ_{dom} are hyperparameters that control the influence of each term.

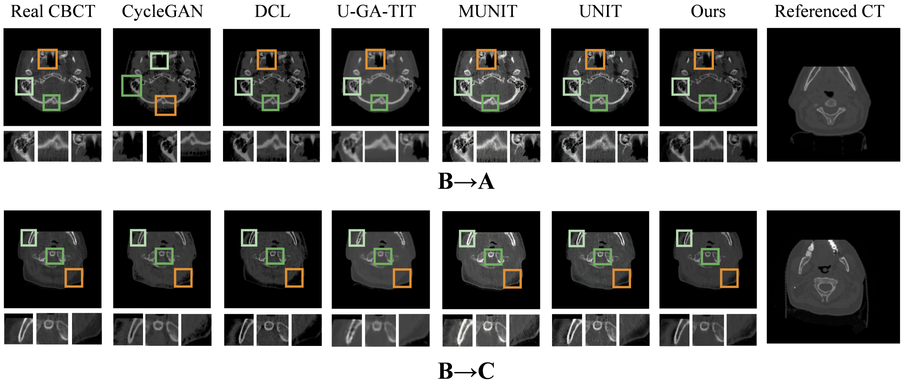

Fig. 4. Visual comparison of CBCT-to-CT translation results using different methods. The top and bottom rows show translations from center B to centers A and C styles respectively, with reference CT images providing style guidance. Zoomed regions highlight local details.

3 Experiments

We utilized a multi-center dataset from the SynthRAD2023 Grand Challenge [22], comprising 180 CBCT-CT sets across three institutions: UMC Utrecht, UMC Groningen, and Radboud Nijmegen. Each institution contributed 60 CT and 60 CBCT volumes, with each 3D volume containing approximately 180-240 slices at resolutions ranging from 224×224 to 256×256 pixels. The original dataset encompassed roughly 66,072 axial slices.

The data preprocessing pipeline involved several steps. Initially, we sliced each volume along the axial (z) dimension to extract 2D images. Subsequently, a mask matching procedure was implemented to ensure anatomical alignment between modalities. Slices with an effective tissue area less than 15% of the total image area were excluded due to their limited diagnostic relevance. The resulting dataset consisted of 64,462 valid slices, which were resized to 256×256 pixels. To create an unpaired training scenario, we performed a global slice-level shuffling across all processed images, deliberately breaking any inherent slice-wise correspondences between modalities. Voxel intensities were then normalized to the range $[-1, 1]$, adhering to standard practice.

3.1 Hyperparameter Settings

We trained the MSDG-StyleNet model using the Adam optimizer with a learning rate of 1×10^{-5}. The batch size was set to 2. The loss function weights were configured as follows: $\lambda_{adv} = 1$, $\lambda_{rec} = 1$, $\lambda_{cyc} = 10$ and $\lambda_{dom} = 1$. The model was trained for 50 epochs. The cycle consistency loss used the L1 loss function, while the GAN loss employed the Hinge loss function. All experiments were conducted on a single GPU with 24GB of memory.

Table 1. Comparison of different models and ablation study results on unpaired image translation metrics.

Network	B→A			B→C		
	FID (↓)	EMD (↓)	1-NN (↓)	FID (↓)	EMD (↓)	1-NN (↓)
CycleGAN [23]	128.49	0.0156	0.9338	130.68	0.0186	0.9308
DCL [24]	125.77	0.0053	0.9706	128.88	0.0123	0.9462
U-GA-TIT [25]	115.32	0.0059	0.9401	110.24	0.0084	0.9325
UNIT [26]	96.49	0.0054	0.7206	90.50	0.0075	0.7771
MUNIT [27]	90.55	0.0113	0.8235	86.61	0.0162	0.7769
w/ Concat (vs. AdaIN)	320.78	0.0191	0.9950	318.42	0.0183	0.9940
w/o Cycle Consistency	88.45	0.0052	0.7306	85.21	0.0073	0.7869
w/o Style Prototypes	92.17	0.0078	0.8088	89.32	0.0089	0.7462
Ours	**81.63**	**0.0042**	**0.7206**	**77.76**	**0.0036**	**0.6615**

3.2 Results and Analysis

We evaluate our method against state-of-the-art unpaired image translation approaches (Fig. 4) and analyze the contribution of individual components through ablation studies.

Our method was quantitatively assessed using three metrics: Fréchet Inception Distance (FID), Earth Mover's Distance (EMD), and 1-Nearest Neighbor Accuracy (1-NN Acc). FID quantifies the similarity between generated and real image distributions, and lower FID scores indicate higher image quality. EMD measures the minimal cost to transform one distribution into another, with lower EMD values suggesting superior distribution matching. 1-NN accuracy assesses the distinguishability between generated and real images, where lower values imply better generation quality, as the discriminator encounters greater difficulty in differentiation.

As demonstrated in Table 1, our method consistently surpasses baseline methods across all metrics in both B→A and B→C translation scenarios. Notably, our method achieves the lowest FID scores (81.63 and 77.76), signifying superior image quality and realism. Traditional methods like CycleGAN [23], DCL [24] and U-GA-TIT [25] exhibit limitations, presenting higher FID and EMD values, suggesting less effective domain adaptation. While UNIT [26] and MUNIT [27] show enhanced performance compared to traditional methods, our approach still yields superior results, particularly in EMD (0.0042 and 0.0036) and 1-NN accuracy metrics.

Ablation study results, presented in the lower section of Table 1, corroborate the effectiveness of key components within our framework. The removal of style prototypes results in a significant performance decline (FID increase of 10.54 and 11.56 points), highlighting their critical role in capturing domain-specific characteristics. The dual consistency mechanism also proves vital, and its absence leads to increased FID scores by 6.82 and 7.45 points, respectively. Moreover, substituting AdaIN with simple feature concatenation degrades performance, further validating the effectiveness of adaptive style modulation in our architecture. These findings collectively demonstrate that each component significantly contributes to achieving robust cross-domain translation while preserving both anatomical consistency and institutional style characteristics.

4 Conclusion

In this study, we present MSDG-StyleNet, a novel framework designed for unsupervised domain-generalized CBCT-to-CT translation, which effectively maintains institution-specific style consistency. Our framework incorporates learnable domain-style prototypes and a dual-level consistency mechanism, enabling robust domain generalization using only single-source training data. The prototype-based style encoding offers a scalable solution for managing diverse institutional styles without escalating model complexity.

Our comprehensive evaluation underscores that MSDG-StyleNet effectively addresses the inherent challenge of maintaining institutional imaging characteristics during CBCT-to-CT translation, which is pivotal for clinical adoption. The framework's success in disentangling content and style, while preserving institutional characteristics, has broader implications for various medical image processing applications. Future research could focus on developing more robust protocols for prototype initialization and adaptation across diverse clinical settings, thereby contributing to the advancement of clinically applicable AI solutions in medical imaging.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. A. C. Miracle and S. K. Mukherji, "Conebeam CT of the head and neck, part 1: physical principles," *AJNR Am. J. Neuroradiol.*, vol. 30, no. 6, pp. 1088–1095, Jun. 2009.
2. S. Kida et al., "Cone Beam Computed Tomography Image Quality Improvement Using a Deep Convolutional Neural Network," *Cureus*, vol. 10, no. 4, p. e2548, Apr. 2018.
3. T. J. Jang, H. S. Yun, C. M. Hyun, J.-E. Kim, S.-H. Lee, and J. K. Seo, "Fully automatic integration of dental CBCT images and full-arch intraoral impressions with stitching error correction via individual tooth segmentation and identification," *Med. Image Anal.*, vol. 93, p. 103096, 2024.
4. D. Hu, Y. Zhang, J. Liu, Y. Zhang, J. L. Coatrieux, and Y. Chen, "PRIOR: Prior-Regularized Iterative Optimization Reconstruction For 4D CBCT," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 11, pp. 5551–5562, 2022.
5. Z. Liang et al., "Leveraging GAN-based CBCT-to-CT translation models for enhanced image quality and accurate photon and proton dose calculation in adaptive radiotherapy," *J. Radiat. Res. Appl. Sci.*, vol. 17, no. 1, p. 100809, 2024.
6. N. Dahiya et al., "Multitask 3D CBCT-to-CT translation and organs-at-risk segmentation using physics-based data augmentation," *Med. Phys.*, vol. 48, no. 9, pp. 5130–5141, 2021.
7. C. Suwanraksa, J. Bridhikitti, T. Liamsuwan, and S. Chaichulee, "CBCT-to-CT Translation Using Registration-Based Generative Adversarial Networks in Patients with Head and Neck Cancer," *Cancers Basel*, vol. 15, no. 7, Mar. 2023.
8. X. Liang, D. Nguyen, and S. B. Jiang, "Generalizability issues with deep learning models in medicine and their potential solutions: illustrated with cone-beam computed tomography (CBCT) to computed tomography (CT) image conversion," *Mach. Learn. Sci. Technol.*, vol. 2, no. 1, p. 015007, Mar. 2021.
9. Y. Pang, Y. Liu, X. Chen, P.-T. Yap, and J. Lian, "SinoSynth: A Physics-Based Domain Randomization Approach for Generalizable CBCT Image Enhancement," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2024, pp. 646–656.
10. A. Alotaibi, "Deep generative adversarial networks for image-to-image translation: A review," *Symmetry*, vol. 12, no. 10, p. 1705, 2020.
11. Y. Pang, J. Lin, T. Qin, and Z. Chen, "Image-to-image translation: Methods and applications," *IEEE Trans. Multimed.*, vol. 24, pp. 3859–3881, 2021.
12. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," arXiv:1505.04597, May 2015.
13. A. Krishna and K. Mueller, "Medical (CT) Image Generation with Style."
14. K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain Generalization: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4396–4415, 2023.
15. L. Zhang et al., "When Unseen Domain Generalization is Unnecessary? Rethinking Data Augmentation," arXiv:1906.03347, Jun. 2019.
16. Z. Su, K. Yao, X. Yang, K. Huang, Q. Wang, and J. Sun, "Rethinking Data Augmentation for Single-Source Domain Generalization in Medical Image Segmentation," *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 2, pp. 2366–2374, Jun. 2023.
17. K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Adversarial Invariant Feature Learning with Accuracy Constraint for Domain Generalization," in *Machine Learning and Knowledge Discovery in Databases*, vol. 11907, Springer International Publishing, 2020, pp. 315–331.

18. A. G. Khoee, Y. Yu, and R. Feldt, “Domain generalization through meta-learning: a survey,” *Artif. Intell. Rev.*, vol. 57, no. 10, p. 285, Sep. 2024.
19. J. S. Yoon, K. Oh, Y. Shin, M. A. Mazurowski, and H.-I. Suk, “Domain Generalization for Medical Image Analysis: A Review,” *Proc. IEEE*, vol. 112, no. 10, pp. 1583–1609, Oct. 2024.
20. Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, “Neural Style Transfer: A Review,” *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 11, pp. 3365–3385, 2020.
21. J. Liu et al., “CBCT-based synthetic CT generation using generative adversarial networks with disentangled representation,” *Quant. Imaging Med. Surg.*, vol. 11, no. 12, pp. 4820–4834, Dec. 2021.
22. E. M. C. Huijben et al., “Generating synthetic computed tomography for radiotherapy: SynthRAD2023 challenge report,” *Med. Image Anal.*, vol. 97, p. 103276, 2024.
23. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.
24. J. Han, M. Shoeiby, L. Petersson, and M. A. Armin, “Dual Contrastive Learning for Unsupervised Image-to-Image Translation,” arXiv:2104.07689, Apr. 2021.
25. J. Kim, M. Kim, H. Kang, and K. Lee, “U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation,” arXiv:1907.10830, Apr. 2020.
26. M.-Y. Liu, T. Breuel, and J. Kautz, “Unsupervised Image-to-Image Translation Networks,” arXiv:1703.00848, Jul. 2018.
27. X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, “Multimodal Unsupervised Image-to-Image Translation,” arXiv:1804.04732, Aug. 2018.