Medical Image Analysis 89 (2023) 102904

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

CDDSA: Contrastive domain disentanglement and style augmentation for generalizable medical image segmentation

Ran Gu [a], Guotai Wang [a,i,*], Jiangshan Lu [a], Jingyang Zhang [b,c], Wenhui Lei [d,i], Yinan Chen [e,g], Wenjun Liao [f], Shichuan Zhang [f], Kang Li [g], Dimitris N. Metaxas [h], Shaoting Zhang [a,e,i]

[a] School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China
[b] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[c] School of Biomedical Engineering, ShanghaiTech University, Shanghai, China
[d] School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China
[e] SenseTime Research, Shanghai, China
[f] Department of Radiation Oncology, Sichuan Cancer Hospital and Institute, University of Electronic Science and Technology of China, Chengdu, China
[g] West China Hospital-SenseTime Joint Lab, West China Biomedical Big Data Center, Sichuan University, Chengdu, China
[h] Department of Computer Science, Rutgers University, Piscataway NJ 08854, USA
[i] Shanghai AI Lab, Shanghai, China

ARTICLE INFO

Keywords:
Disentanglement
Domain generalization
Contrastive learning
Medical image segmentation

ABSTRACT

Generalization to previously unseen images with potential domain shifts is essential for clinically applicable medical image segmentation. Disentangling domain-specific and domain-invariant features is key for Domain Generalization (DG). However, existing DG methods struggle to achieve effective disentanglement. To address this problem, we propose an efficient framework called Contrastive Domain Disentanglement and Style Augmentation (CDDSA) for generalizable medical image segmentation. First, a disentangle network decomposes the image into domain-invariant anatomical representation and domain-specific style code, where the former is sent for further segmentation that is not affected by domain shift, and the disentanglement is regularized by a decoder that combines the anatomical representation and style code to reconstruct the original image. Second, to achieve better disentanglement, a contrastive loss is proposed to encourage the style codes from the same domain and different domains to be compact and divergent, respectively. Finally, to further improve generalizability, we propose a style augmentation strategy to synthesize images with various unseen styles in real time while maintaining anatomical information. Comprehensive experiments on a public multi-site fundus image dataset and an in-house multi-site Nasopharyngeal Carcinoma Magnetic Resonance Image (NPC-MRI) dataset show that the proposed CDDSA achieved remarkable generalizability across different domains, and it outperformed several state-of-the-art methods in generalizable segmentation. Code is available at https://github.com/HiLab-git/DAG4MIA.

1. Introduction

Deep learning with Convolutional Neural Networks (CNNs) has achieved remarkable performance in medical image segmentation (Ronneberger et al., 2015; Shen et al., 2017; Gu et al., 2021), and most existing models are built on the assumption that training and testing images are from the same domain and have very similar, if not the same, distributions. However, in clinical practice, this assumption does often not hold due to several factors such as the differences in scanning devices, imaging protocols, patient groups and image quality between training and testing images, where the testing images are usually acquired from a different medical centre than the training set. Such differences (a.k.a domain shift) can substantially degrade the model's performance at test time (Ganin et al., 2016; Kamnitsas et al., 2017; Guan and Liu, 2021).

To address this problem, many Domain Adaptation (DA) methods have been explored to transfer knowledge from a set of labelled images in a source domain to images in a target domain (Tzeng et al., 2017; Wu et al., 2022; Gu et al., 2022). However, the DA methods need to tune the model's parameters based on a set of images in the target domain, which is not only time-consuming but also impractical if the target domain is not known in advance (Chen et al., 2018). Moreover, the model needs to be adapted to each target domain respectively and

* Corresponding author at: School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China.
E-mail addresses: Guotai.Wang@uestc.edu.cn (G. Wang), Zhangshaoting@uestc.edu.cn (S. Zhang).

https://doi.org/10.1016/j.media.2023.102904
Received 10 November 2022; Received in revised form 6 June 2023; Accepted 12 July 2023
Available online 18 July 2023
1361-8415/© 2023 Published by Elsevier B.V.

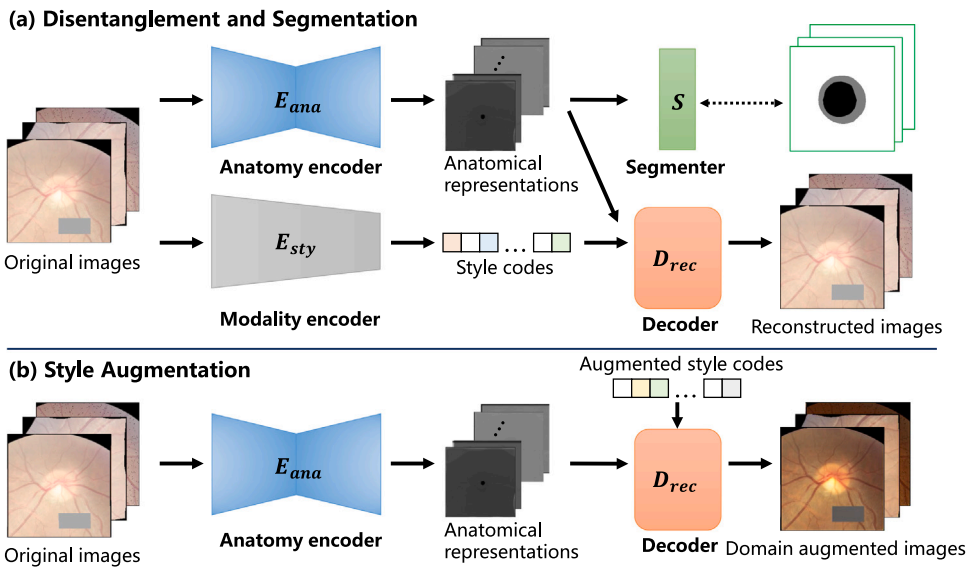

Fig. 1. Workflow of our proposed Contrastive Domain Disentanglement and Style Augmentation (CDDSA) method. (a) shows the disentanglement and segmentation networks, where an anatomy encoder and a style encoder obtain anatomical representations and style codes respectively, and they are regularized by a decoder to reconstruct the input image. The segmentor takes domain-invariant anatomical representations as input to obtain the segmentation results. (b) represents the style augmentation strategy, where we combine anatomical representations from a given image with augmented style codes to generate images in a new domain.

is faced with the problem of catastrophic forgetting on previous domains, which is not scalable when applied to a range of new unknown domains.

In contrast to DA, Domain Generalization (DG) which encourages a model to be generalizable to unseen domains is more appealing and efficient as it does not need to tune the model after training. Recently, domain generalization has attracted increasing attention in the field of computer vision (Wang et al., 2022) and medical image analysis (Wang et al., 2023). Existing DG methods mainly include image- and feature-based approaches. For image-based approaches, data augmentation has been widely used to improve the model generalization performance (Zhang et al., 2020). In contrast, feature-based methods mainly focus on representation learning to extract the most representative features for better generalization across domains (Wang et al., 2020; Gu et al., 2021). However, both image- and feature-based methods are limited by the capacity and representation power of the knowledge pool and have a limited ability to recognize sufficient invariant features across different domains.

Recently, disentanglement has been introduced to computer vision that aims to explicitly decompose features into domain-invariant contents and domain-specific styles (Tran et al., 2017). It has also been employed to learn domain-invariant features for domain adaptation on multi-modality medical image segmentation datasets. Yang et al. (2019) applied disentangled representations to unsupervised domain adaptation for liver segmentation. They decomposed the images from two domains into a shared domain-invariant content space and a domain-specific style space and used representations in the content space for segmentation. Pei et al. (2021) used disentangled domain-invariant and domain-specific features for cardiac image segmentation across two modalities, and introduced a zero-loss to enhance the disentanglement. However, most existing disentangling methods are based on Generative Adversarial Networks (GANs), where a content encoder and a style encoder need to be trained for each known modality/domain, and multiple discriminators are involved, leading to a complex training process. Although GAN-based disentanglement methods are suitable for domain adaptation, they face challenges when it comes to domain generalization. This is because they usually require domain-specific content/style encoders, reconstruction decoders and discriminators, and the number of required encoders, decoders and discriminators may increase as the number of domains grows, making the training process very complex. What is more, these methods cannot be directly applied to unseen target domains since they need data from each domain for training. Consequently, GAN-based disentanglement methods are not optimal solutions for domain generalization problems.

In this work, we propose a novel GAN-free disentanglement framework named Contrastive Domain Disentanglement and Style Augmentation (CDDSA) for domain-generalizable medical image segmentation. As shown in Fig. 1, it decomposes medical images in different domains into domain-invariant anatomical representations and domain-specific style codes with only one pair of anatomy encoder and style encoder, which is regularized by a decoder that accepts an anatomical representation and a style code to reconstruct an image. The encoders and decoder are shared across different domains, without adversarial learning and domain-specific training, which is efficient and scalable to multiple domains. Our method was inspired by Spatial Decomposition Network (SDNet) (Chartsias et al., 2019) that implements feature disentanglement without GAN. Note that SDNet (Chartsias et al., 2019) was proposed for semi-supervised learning, modality transformation and multi-modal image segmentation. It can only perform disentanglement and image reconstruction on seen domains with poor generalizability in unseen domains. The main reason is that SDNet lacks effective constraints on the style codes to encourage them to be domain-specific, which limits the ability to extract domain-invariant feature representations. In addition, it restricts the anatomical representations as binary codes, leading to a limited representation ability for effective image reconstruction.

Differently from SDNet (Chartsias et al., 2019), our CDDSA is proposed for domain-generalizable segmentation of medical images. To improve the disentanglement performance, we relax the anatomical representation to soft values and propose a domain style contrastive learning loss to encourage the style codes in different domains to be discriminative from each other, which improves the model's ability to recognize domain-invariant anatomical representations that are sent to a segmentor to obtain segmentation results. As the segmentor is not affected by domain-specific features, it has a high generalizability across different domains. In addition, based on the extracted style codes in training domains, we can generate a new random style code and combine it with an existing anatomical representation to simulate images in an unseen domain with new styles using the decoder, i.e., style augmentation, which further improves the generalizability of our framework.

To the best of our knowledge, this is the first work in the literature to propose feature disentanglement learning for domain-generalizable medical image segmentation. The contributions of our method are summarized as follows:

(1) We introduce a novel framework CDDSA using GAN-free disentanglement for domain generalization in medical image segmentation. It achieves generalizability by segmentation from decomposed domain-invariant representations extracted by a single anatomy Encoder that is shared across domains and more efficient and scalable than GAN-based disentanglement.
(2) To make the disentangled domain-specific style codes more representative and distinguishable, we propose domain style contrastive learning, which forces the style codes from the same domain and different domains to be similar and dissimilar, respectively.
(3) We propose style augmentation based on the disentangled anatomical representations and style codes to simulate images from unseen domains with different styles, which further improves the generalizability of the disentanglement and segmentation models.

Comprehensive experimental results on multi-domain fundus images and multi-domain nasopharyngeal carcinoma magnetic resonance images (NPC-MRI) showed that our proposed CDDSA achieved high generalization on unseen domains, and it outperformed several state-of-the-art domain generalization methods.

2. Related works

2.1. Domain generalization for medical image analysis

Recently, domain generalization has attracted increasing attention to avoid dramatic performance degradation when inferring with images from unseen domains (Li et al., 2018b). It aims to learn a model from single or multiple source domains to make it directly applicable for unseen target domains without extra training (Li et al., 2018a; Dou et al., 2019; Wang et al., 2022). Existing DG methods mainly include meta-learning methods, data-based methods and feature-based methods. Meta-learning (Liu et al., 2020; Dou et al., 2019) splits a set of source domains into meta-train and meta-test subsets and adopts meta-optimization that iteratively updates model parameters to improve performance on the meta-test subset to simulate the situation when inferring on unseen domains. Liu et al. (2021) combined meta-learning with federated learning to achieve privacy-preserving generalizable segmentation through continuous frequency space interpolation across clients. However, the meta-optimization process is highly time-consuming since all potential splitting results of meta-train and meta-test should be considered during training (Dou et al., 2019).

Data-based approaches usually use different data augmentation strategies for improving the model's generalizability. Zhang et al. (2020) a deep stacked transformation assuming that the shift between different domains can be simulated by extensive data augmentation on a single domain. Fick et al. (2021) utilized Cycle-GAN (Zhu et al., 2017) to transform images from one certain domain to other domains for augmentation. Li et al. (2022) proposed Mixed Task Sampling (MTS) to enhance the variety of task-level training samples. Mixup in frequency domains (Liu et al., 2021; Zhou et al., 2022a) has also been used to synthesize new images for model generalization. However, the efficiency of data augmentation largely depends on the ability to cover the data distribution in unseen domains, hence requiring empirical settings and even data-specific modifications.

Feature-based approaches use domain-adaptive feature calibration or learn domain-invariant features to deal with domain generalization (Wang et al., 2020; Muandet et al., 2013; Li et al., 2018c). Wang et al. (2020) introduced a domain-oriented feature embedding framework that dynamically updates the domain-specific prior knowledge to make the semantic features more discriminative. Hu et al. (2022b) proposed a dynamic convolutional head to make the model's convolutional parameters adaptive to unseen target domains. Gu et al. (2021) proposed a domain composition and attention method that calibrates the input feature based on attention coefficients represented by a representation bank. However, these methods did not explicitly obtain domain-invariant features for domain generalization, and they did not separate features into purely domain-specific and domain-invariant representations well, leading to a limited performance on domain generalization.

2.2. Disentanglement representation learning

Disentanglement explicitly decomposes features into domain-invariant contents and domain-specific styles (Bengio et al., 2013; Gatys et al., 2016). In addition to applications such as image synthesis (Chartsias et al., 2019), artefact removal and multi-task learning (Meng et al., 2019) for medical image analysis, it is widely adopted for domain adaptation (Yang et al., 2019; Pei et al., 2021). Yang et al. (2019) used disentanglement to obtain domain-invariant content features for liver segmentation with domain adaptation. Xie et al. (2020) used disentanglement to improve the performance of image translation for domain adaptation, and they disentangled the content features from domain information for both the source and translated images. Pei et al. (2021) applied disentanglement-based domain adaptation for cardiac image segmentation, and introduced a zero loss to enhance disentanglement. Ning et al. (2021) proposed a bidirectional unsupervised DA framework based on disentangled representation learning for equally competent two-way DA performances on cardiac image segmentation. Despite their good performance on DA, these works achieve disentanglement based on GANs, where multiple discriminators are needed in the adversarial training process that is complex and tricky to optimize. What is more, they need to have access to images for target domains during training and are not applicable to DG tasks that involve unseen domains. Chartsias et al. (2019) proposed a GAN-free Spatial Decomposition Network (SDNet) that decomposes an input image into a spatial factor (anatomy) and a non-spatial factor (style), and applied it to semi-supervised segmentation and image synthesis. However, it performs disentanglement and reconstruction well only on seen domains and can hardly deal with unseen domains that are not involved in training.

2.3. Contrastive learning

Contrastive learning is a self-supervised learning method to learn feature representations by enforcing positive pairs to be close and negative pairs to be distant (Hadsell et al., 2006). Previous contrastive learning methods were mainly proposed to pre-train a powerful and representational feature extractor that can distinguish similar and dissimilar samples (He et al., 2020; Chen et al., 2020). For computer vision and medical image analysis, contrastive learning has been mainly used for annotation-efficient learning. For example, Kang et al. (2019) proposed a contrastive adaptation network that minimizes the intra-class domain discrepancy and maximizes the inter-class domain discrepancy. Chaitanya et al. (2020) used contrastive learning of global and local features sequentially for 3D medical image segmentation with limited annotations. Lei et al. (2021) proposed contrastive learning of relative position regression for one-shot object localization in 3D medical images. You et al. (2022) proposed a contrastive voxel-wise representation learning to effectively learn low-level and high-level features for semi-supervised medical image segmentation. Unlike these works, we design a contrastive learning strategy to enhance disentanglement between domain-invariant and domain-specific features to deal with domain generalization problems.

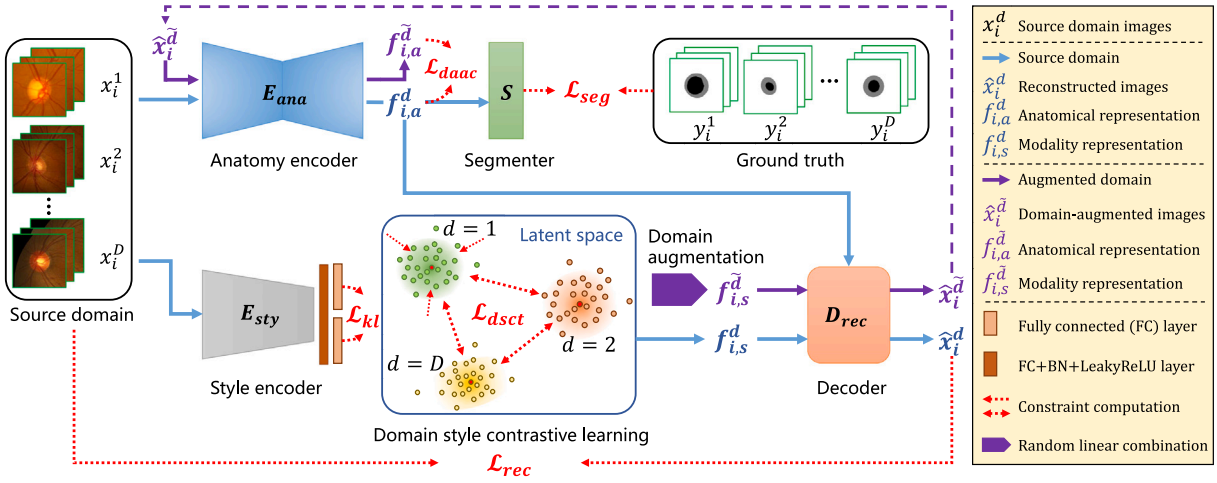

Fig. 2. Overview of the proposed Contrastive Domain Disentanglement and Style Augmentation (CDDSA) network for multi-domain generalizable segmentation. We use an anatomy encoder E_{ana} and a modality encoder E_{sty} to extract anatomical representations f_a^d and style codes f_s^d, respectively. A reconstruction decoder D_{rec} takes f_a^d and f_s^d as input and obtains a reconstructed image $\hat{x}^d$. The decomposed anatomical representations f_a^d is further used for segmentation. $\hat{x}^{\tilde{d}}$ is our simulated image using domain-augmentation strategy. $\hat{x}^{\tilde{d}}$ will further input into E_{ana} to get its anatomical representation $f_a^{\tilde{d}}$. $\mathcal{L}_{saac}$ is used to encourage the consistency between f_a^d and $f_a^{\tilde{d}}$.

3. Methods

For the domain generalization problem, the training set consists of images from D domains and can be denoted as $\mathcal{D} = \{(x_i^d, y_i^d)\}_{i=1}^{N_d}$ $(d = 1, 2, \ldots, D)$, where x_i^d depicts the ith training sample from the dth source domain with its corresponding ground-truth annotation y_i^d. N_d denotes the number of training samples in domain d.

Our proposed Contrastive Domain Disentanglement and Style Augmentation (CDDSA) framework is illustrated in Fig. 2. Firstly, we employ a disentangle network containing an anatomy encoder E_{ana} and a style encoder E_{sty} to decompose an image into a domain-invariant anatomical representation and a domain-specific modality representation (i.e., style code), and they can be used to reconstruct the input image based on a decoder D_{rec}. We further send the disentangled anatomical representation into a segmentor S to predict the segmentation mask. Secondly, to boost the disentanglement performance with more discriminative style codes across different domains, we introduce domain style contrastive learning that forces the decomposed modality representations to have low intra-domain discrepancy and high inter-domain discrepancy. Thirdly, to further enhance model generalization, we proposed a style augmentation strategy to randomly generate style codes and combine them with given anatomical representations to reconstruct images with new styles that are not present in the training set.

3.1. Domain disentangle network

As shown in Fig. 2, for an input image x_i^d, we send it to an anatomy encoder E_{ana} and a style encoder E_{sty} to obtain an anatomical representation $f_{i,a}^d$ and a modality representation (style code) $f_{i,s}^d$, respectively. Then $f_{i,a}^d$ and $f_{i,s}^d$ are sent to a decoder D_{rec} to reconstruct an input-like images $\hat{x}_i^d$, and a reconstruction loss $\mathcal{L}_{rec}$ is used to encourage the consistency between x_i^d and $\hat{x}_i^d$. A segmentor S takes $f_{i,a}^d$ as input to obtain the segmentation result.

3.1.1. Anatomy encoder and segmenter

To decompose domain-invariant anatomical representations, we employ a U-Net as the backbone to implement E_{ana}. We modify U-Net by setting the output channel of the last layer as T and using tanh as the activation function in that layer. Let H and W represent the height and width of the input image x_i^d respectively, the output of E_{ana} is denoted as $f_{i,a}^d \in [-1, 1]^{H \times W \times T}$, and we assume that each channel of $f_{i,a}^d$ emphasizes some anatomical information. Differently from SD-Net (Chartsias et al., 2019) that constrains $f_{i,a}^d$ to take binary values that may lose many details of object structures, we aim to reserve enough structural information for accurate image reconstruction and further style augmentation, and therefore soften the anatomical representation with a $tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ activation function in the last layer. The anatomical representation extraction procedure is formulated as:

$$f_{i,a}^d = E_{ana}(x_i^d) \tag{1}$$

Then, the decomposed anatomical representation $f_{i,a}^d$ is fed into a segmentation network S to obtain a segmentation probability map $p_i^d = S(f_{i,a}^d)$. Let y_i^d denote the ground truth, and the supervised segmentation loss for domain d is:

$$\mathcal{L}_{seg} = \frac{1}{2N_d} \sum_{i=1}^{N_d} \left(\mathcal{L}_{Dice}(p_i^d, y_i^d) + \mathcal{L}_{ce}(p_i^d, y_i^d) \right) \tag{2}$$

where we use a hybrid segmentation loss that consists of a Dice loss $\mathcal{L}_{Dice}$ and a cross-entropy loss $\mathcal{L}_{ce}$.

3.1.2. Style encoder

The domain-specific modality representations are obtained by a style encoder E_{sty} that is implemented by a Variational Autoencoder (VAE) (Kingma and Welling, 2014). The VAE learns a low dimensional latent space so that the learned latent representations match a prior distribution of an isotropic multivariate Gaussian $p(z) = \mathcal{N}(0, 1)$. Given the input x_i^d, E_{sty} predicts the mean u_i^d and variance v_i^d of the distribution of a latent code $z \in \mathbb{R}^{1 \times Z}$, where Z is the length of the latent code. The style code $f_{i,s}^d$ of an input image $x_{i,s}^d$ is sampled from the distribution characterized by mean u_i^d and variance v_i^d. The VAE is trained to minimize a reparameterization error, and a KL divergence loss is computed between the estimated Gaussian distribution $q(z|u_i^d, v_i^d)$ and the unit Gaussian $p(z)$:

$$\mathcal{L}_{kl} = D_{kl}\left(q(z|u_i^d, v_i^d) \parallel p(z)\right) \tag{3}$$

where $D_{kl}(p \parallel q) = \sum p(x) log \frac{p(x)}{q(x)}$. When training is finished, sampling a vector from the unit Gaussian over a latent space can obtain a new style code, and we send it together with an anatomical representation to the decoder to obtain a reconstructed image, where the decoder is used as a generative model, as detailed in the following.

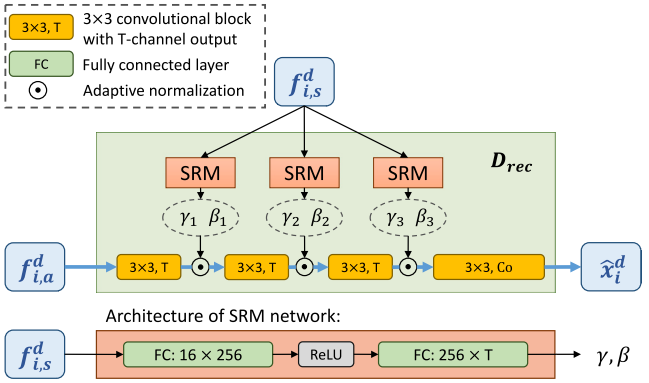

Fig. 3. Framework of the reconstruction decoder D_{rec}. T is the channel of feature maps and Co represents the channel of output reconstructed image.

3.1.3. Reconstruction decoder

Fig. 3 shows the structure of our reconstruction decoder D_{rec} to generate an image $\hat{x}_i^d$ given two decomposed representations $f_{i,a}^d$ and $f_{i,s}^d$. The collaboration of the two representations acts as a repainting mechanism where the anatomical representation $f_{i,a}^d$ is used to derive the anatomical content, and the modality representation $f_{i,s}^d$ is used to colour the style distribution on the whole image (Huang et al., 2018).

Specifically, the decoder uses four convolutional blocks to map $f_{i,a}^d$ to a reconstructed image conditioned on three Style Reconstruction Modules (SRM), as shown in Fig. 3. For the intermediate feature map obtained by each convolutional block in the decoder, we apply Adaptive Instance Normalization (AdaIN) to control the output style, where the affine transformation parameters (scale and bias) are predicted by an SRM that takes $f_{i,s}^d$ as input. Let $F_{i,c}$ represent the cth channel of the intermediate feature map, we use two Fully Connected (FC) layers with a ReLU activation to implement the SRM that maps the modality representation $f_{i,s}^d$ to the scale $\gamma_{i,c}$ and bias $\beta_{i,c}$ that are used by affine transformation of AdaIN:

$$AdaIN(F_{i,c}|\gamma_{i,c}, \beta_{i,c}) = \gamma_{i,c}\frac{F_{i,c} - \mu(F_{i,c})}{\sigma(F_{i,c})} + \beta_{i,c} \tag{4}$$

where each channel of the intermediate feature map is normalized separately, and we apply AdaIN with SRM after each of three convolutional blocks in the decoder respectively, as shown in Fig. 3. By mapping $f_{i,s}^d$ to the scale and bias values for each intermediate feature map, the reconstruction decoder D_{rec} adaptively repaints style distribution on the anatomical representation $f_{i,a}^d$ in a coarse-to-fine manner. We use $\hat{x}_i^d$ to denoted the reconstructed image based on $f_{i,s}^d$ and $f_{i,a}^d$, and it is obtained by:

$$\hat{x}_i^d = D_{rec}(f_{i,s}^d, f_{i,a}^d) \tag{5}$$

As $f_{i,s}^d$ and $f_{i,a}^d$ are obtained from x_i^d, the reconstructed image $\hat{x}_i^d$ should be as close as possible to x_i^d. Therefore, a reconstruction loss is employed to train the anatomy encoder E_{ana}, style encoder E_{sty} and reconstruction decoder D_{rec}:

$$\mathcal{L}_{rec} = \frac{1}{N_d}\sum_{i=1}^{N_d}\left|x_i^d - \hat{x}_i^d\right| \tag{6}$$

where we simply define the reconstruction loss as the Mean Absolute Error (MAE) loss due to its robustness to outliers.

3.2. Domain style contrastive learning

An effective disentanglement expects that the style code $f_{i,s}^d$ is domain-specific, but the reconstruction loss $\mathcal{L}_{rec}$ does not provide sufficient supervision for achieving domain-specific style codes. To address the problem and make the model decompose more discriminative modality representations for different domains, we propose a domain style contrastive learning strategy to explicitly constrain the disentangled style code $f_{i,s}^d$.

Let x_i^d and x_j^d represent two different samples from the same domain d in the training set, and their style codes obtained by E_{sty} are denoted as $f_{i,s}^d$ and $f_{j,s}^d$, respectively. We define $(f_{i,s}^d, f_{j,s}^d)$ as a positive pair for maximizing their similarity. At the same time, for N samples each from a different domain d' ($d' \in [0, 1, \dots, D]$ and $d' \neq d$), their corresponding style codes compose a negative set $\mathcal{N}_i^d$ for $f_{i,s}^d$, and each element in $\mathcal{N}_i^d$ should have a minimized similarity compared with $f_{i,s}^d$. Following the standard formula of self-supervised contrastive loss InfoNCE (Oord et al., 2018; Wang et al., 2021), we define our domain style contrastive loss as:

$$\mathcal{L}_{dsct} = -log\frac{e^{sim(f_{i,s}^d, f_{j,s}^d)/\tau}}{e^{sim(f_{i,s}^d, f_{j,s}^d)/\tau} + \sum_{f\in\mathcal{N}_i^d} e^{sim(f_{i,s}^d, f)/\tau}} \tag{7}$$

where $sim(\cdot,\cdot)$ is the cosine similarity, and $\tau = 0.1$ is the temperature scaling parameter following the settings in the classical contrastive learning (Chen et al., 2020). In practice, to save the computational cost during training, we fetch b samples for each domain in a mini-batch, and their style codes are saved in a list. We fetch b samples for each domain to compose a mini-batch of size bD, and their style codes are saved in a list. For $x_i^d (i \leqslant b)$ in the dth domain, we randomly select another sample x_j^d (j ⩽ b and j ≠ i) from the same domain in the mini-batch, and use (x_i^d, x_j^d) as a positive pair. Therefore, b positive pairs are considered for each domain in a mini-batch. All the $b(D-1)$ samples from the other domains in the mini-batch are set as the corresponding negative samples of x_i^d.

3.3. Style augmentation with anatomical consistency

Based on the disentangled anatomical representations, style codes and the decoder, we can augment the style of an image by replacing its style code during image reconstruction, and therefore propose a style augmentation strategy to automatically generate images in new domains with different styles. At each iteration of training, we denote the style codes of a batch as a style code bank $\mathcal{F} = \{f_{i,s}^d | i = 1, 2, \dots, B; d = 1, 2, \dots, D\}$, where the batch has B samples for each domain. Based on the style codes in $\mathcal{F}$, we obtain a new style code using a linear combination of them with random weights:

$$f_s^{\tilde{d}} = \sum_{i=1}^{|\mathcal{F}|}\alpha_i \mathcal{F}_i \tag{8}$$

where $f_s^{\tilde{d}}$ is a generated style code that is assumed to be from an unseen domain $\tilde{d}$. $\mathcal{F}_i$ is the ith element in the style code bank, and the weight $\alpha_i \in [-1, 1]$ is randomly sampled from a uniform distribution.

Given an anatomical representation $f_{i,a}^d$ from an image in the source domain, we repaint it with the new style code $f_s^{\tilde{d}}$ to generate a new image $\hat{x}_i^{\tilde{d}}$:

$$\hat{x}_i^{\tilde{d}} = D_{rec}(f_s^{\tilde{d}}, f_{i,a}^d) \tag{9}$$

Since the generated image $\hat{x}_i^{\tilde{d}}$ and the real image x_i^d share the same anatomical representation $f_{i,a}^d$, we introduce an anatomical consistency loss $\mathcal{L}_{saac}$ that forces the anatomy encoder E_{ana} to obtain domain-invariant anatomical representations in spite of the different styles between $\hat{x}_i^{\tilde{d}}$ and x_i^d:

$$\mathcal{L}_{saac} = \frac{1}{N_d}\sum_{i=1}^{N_d}\left|f_{i,a}^d - E_{ana}(\hat{x}_i^{\tilde{d}})\right| \tag{10}$$

where the MAE loss is used for anatomical consistency.

Table 1
Statistics of retinal fundus images in four domains used in our experiment following Wang et al. (2020).

Domain No.	Dataset	Cases (train/test)	Scanner
Domain 1	Drishti-GS	50/51	Aravind eye hospital
Domain 2	RIM-ONE-r3	99/60	Nidek AFC-210
Domain 3	REFUGE-train	320/80	Zeiss Visucam 500
Domain 4	REFUGE-val	320/80	Canon CR-2

3.4. Overall loss

As a summary of the proposed CDDSA framework, the overall loss function for training is formulated as:

$$\mathcal{L} = \mathcal{L}_{seg} + \lambda_1 \mathcal{L}_{kl} + \lambda_2 \mathcal{L}_{rec} + \lambda_3 \mathcal{L}_{dsct} + \lambda_4 \mathcal{L}_{saac} \tag{11}$$

where $\mathcal{L}_{seg}$ is the supervised segmentation loss (Eq. (2)), $\mathcal{L}_{kl}$ is the KL divergence loss for style encoder (Eq. (3)), $\mathcal{L}_{rec}$ is the image reconstruction loss (Eq. (6)). $\mathcal{L}_{dsct}$ and $\mathcal{L}_{saac}$ are the domain style contrastive loss (Eq. (7)) and style augmentation-based anatomical consistency loss (Eq. (10)), respectively. $\lambda_1, \lambda_2, \lambda_3$ and λ_4 act as trade-off parameters for different loss terms.

4. Experiments and results

4.1. Datasets and implementation details

In this study, we evaluated our proposed CDDSA and compared it with several state-of-the-art DG methods on a public multi-domain fundus image dataset and an in-house multi-domain nasopharyngeal carcinoma MRI dataset.

Multi-domain Fundus Image Dataset: For a fair comparison with state-of-the-art DG methods, we evaluated our approach for Optic Cup (OC) and Disc (OD) segmentation on a public multi-domain retinal fundus image dataset[1] (Wang et al., 2020). The dataset was collected from four public fundus image datasets obtained by different scanners at different sites that have distinct domain discrepancies in visual appearance and image quality: Domain 1 is from the Drishti-GS (Sivaswamy et al., 2015) dataset containing 50 and 51 images for training and testing, respectively; Domain 2 is from the RIM-ONE (Fumero et al., 2011) dataset containing 99 and 60 images for training and testing, respectively; and the Domain 3 and 4 are from REFUGE (Orlando et al., 2020) challenge's training and validation datasets, respectively, and both of them contain 320 and 80 images for training and testing. These images were centre-cropped to the size of 384 × 384 during preprocessing.

To evaluate the generalizability of OC/OD segmentation models in unseen domains, we followed the leave-one-domain-out cross validation strategy in DoFE (Wang et al., 2020), where each time three domains were used for training and the other domain was used as the unseen testing domain. In total, there are 789 and 271 images for training and testing, respectively. The statistics of these multi-domain retinal fundus images are summarized in Table 1. In training, we followed DoFE (Wang et al., 2020) to adopt a series of basic data augmentations including random cropping to the size of 256 × 256, rotating, flipping, noising and brightness augmentation etc, to enhance the diversity of training samples for a fair comparison.

Multi-domain Nasopharyngeal Carcinoma MRI Dataset: We collected an in-house multi-domain Nasopharyngeal Carcinoma (NPC) MRI dataset for nasopharynx Gross Tumor Volume (GTVnx) segmentation. It was collected from two hospitals with four different imaging protocols (Liao et al., 2022) (i.e., four domains): T1-weighted imaging, gadolinium contrast-enhanced T1-weighted (CE-T1) imaging, T1 water imaging and T2 water imaging, respectively. Images in Domain 1 were collected from Sichuan Provincial People's Hospital (SPPH) with a slice thickness of 6–7.75 mm, and images in Domain 2–4 were collected from West China Hospital (WCH) with a slice thickness of 3 mm. In total, there were 189 volumes each from a specific patient, and they were split into 114 for training and 75 for testing. The corresponding slice numbers for training and testing were 1427 and 994, respectively. The volume and slice numbers for each domain are detailed in Table 2.

For preprocessing, we unified the orientation of different volumes into the standard RAI (right to left, anterior to posterior, inferior to superior in the x-, y-, and z-axes, respectively). The voxel intensity was clipped by the 0.1 and 99.9 percentiles of each volume and then normalized to [0, 255]. Each volume was firstly cropped along the z-axis based on the slices containing GTVnx delineation and then centre-cropped with a 256 × 256 window in the x-y plane. We used 2D networks for the GTVnx segmentation in each slice and stacked the results into a 3D volume for evaluation, and a leave-one-domain-out cross validation strategy was also employed during the experiment.

Implementation Details: Training and inference were implemented on one NVIDIA GeForce GTX 1080 Ti GPU. The anatomical representation E_{ana} was implemented by U-Net as the backbone, with channel numbers of 16, 32, 64, 128 and 256 at five resolution scales, respectively. We set the channel number of anatomical representations as $T = 8$. The segmentor S consists of two convolutional blocks. The first block has a convolution layer with a kernel size of 3 × 3 followed by BN and LeakyReLU (sloop = 0.2), and the second block has a 1 × 1 convolution layer followed by Softmax to obtain a segmentation probability map. The style encoder E_{sty} has convolutional blocks each with a down-sampling layer to reduce the resolution, and the output of the last convolutional block is sent to two fully connected layers to obtain the mean and variance of a Gaussian distribution for the latent style code, and the size of the latent style code was set as $Z = 16$.

For the weights in the total loss function, we first follow the disentangle baseline SDNet (Chartsias et al., 2019) to set λ_1 and λ_2 to 1.0 and 0.001, respectively. As the scales of $\mathcal{L}_{dsct}$ and $\mathcal{L}_{saac}$ were around 10.0 and 1.0 times of $\mathcal{L}_{seg}$, we set λ_3 and λ_4 to 0.1 and 1.0 respectively to balance the scales of the loss terms. The networks were trained with the Adam optimizer, and the learning rate was initialized to 10^{-3} and decayed to 95% when the performance did not improve in 8 epochs. In a mini-batch, the image/slice number for each domain was 8 and 6 for the fundus image and NPC-MRI datasets, respectively. The training epoch was 400 for the fundus image and NPC-MRI datasets. For a fair comparison, all the methods were trained from scratch without using the pre-trained weights. The trained models are deployed at the SenseCare platform (Duan et al., 2020) to support clinic research. To measure the segmentation performance quantitatively, we adopt the Dice score (Dice) and Average Symmetric Surface Distance (ASSD) for evaluation.

4.2. Fundus image segmentation

4.2.1. Comparison with state-of-the-art DG methods

For the domain generalization study, we conducted leave-one-domain-out cross validation on the multi-domain fundus image dataset. We first considered all the available training domains as a single dataset (i.e., ignoring the domain shift in the training set) and trained a U-Net (Ronneberger et al., 2015) using a standard Dice loss, and directly applied it to the unseen domain, which is referred to as '**Inter-domain**' and serves as a lower bound of the experiment. Then, for each domain, we trained and tested the U-Net with the training and testing sets respectively, i.e., no unseen domain involved, which serves as the upper bound for DG and is referred as '**Intra-domain**'. For DG methods, we compared our proposed CDDSA with four representative state-of-the-art approaches: BigAug (Li et al., 2018b) based on data augmentation, DoFE (Wang et al., 2020) based on domain-oriented feature embedding,

[1] https://github.com/emma-sjwang/Dofe.

Table 2
Statistics of the in-house multi-domain nasopharyngeal carcinoma MRI dataset (Liao et al., 2022).

Domain No.	Sequence	Slice thickness (mm)	Volumes (train/test)	Slices (train/test)	Scanner
Domain 1	T1	6–7.75	39/26	305/201	SPPH - Siemens
Domain 2	CE-T1	3	27/18	359/234	WCH - Siemens
Domain 3	T1-water	3	24/15	402/302	WCH - Siemens
Domain 4	T2-water	3	24/16	361/257	WCH - Siemens

Table 3
Comparison of Dice (%) by different DG methods on the multi-site fundus image dataset. CDDSA⋄ means that the new style code for style augmentation was randomly sampled from a Gaussian distribution rather than obtained by a random linear combination of style codes in the source domains.

Methods	Domain 1		Domain 2		Domain 3		Domain 4		Avg	
	Cup	Disc	Cup	Disc	Cup	Disc	Cup	Disc	Cup	Disc
Lower bound (Inter-domain)	74.38 ± 12.96	96.67 ± 2.04	77.71 ± 20.84	85.05 ± 14.67	79.72 ± 9.51	90.01 ± 5.81	86.63 ± 8.52	89.55 ± 3.26	79.61	90.32
Upper bound (Intra-domain)	83.35 ± 13.99	96.10 ± 1.88	81.53 ± 9.42	94.62 ± 3.01	87.57 ± 7.59	95.91 ± 1.85	88.88 ± 7.10	95.58 ± 1.98	85.33	95.55
BigAug (Zhang et al., 2020)	82.36 ± 11.74	93.73 ± 9.29	75.45 ± 15.01	87.83 ± 11.17	84.32 ± 9.45	91.99 ± 10.72	85.32 ± 7.50	92.97 ± 6.58	81.86	91.63
DoFE (Wang et al., 2020)	80.25 ± 10.84	95.61 ± 1.45	78.97 ± 14.80	88.74 ± 4.58	84.81 ± 7.71	92.81 ± 2.63	86.65 ± 6.39	93.46 ± 2.43	82.67	92.66
FedDG (Liu et al., 2021)	79.84 ± 13.55	93.50 ± 4.11	76.57 ± 13.95	88.74 ± 4.91	84.23 ± 6.80	93.73 ± 3.22	85.33 ± 10.19	94.03 ± 4.14	81.49	92.50
DCA-Net (Gu et al., 2021)	82.16 ± 12.23	94.39 ± 2.94	80.63 ± 15.58	**91.50 ± 2.78**	84.48 ± 7.77	91.63 ± 4.38	**87.11 ± 12.67**	93.05 ± 4.98	83.60	92.64
Baseline	80.63 ± 11.55	95.02 ± 2.65	79.35 ± 13.66	89.76 ± 3.20	83.29 ± 8.04	**93.67 ± 3.36**	84.12 ± 11.33	93.51 ± 4.03	81.85	92.99
$+\mathcal{L}_{dsct}$	80.64 ± 11.94	96.11 ± 2.95	80.13 ± 17.39	88.27 ± 1.23	85.20 ± 8.06	92.36 ± 2.59	86.33 ± 9.12	93.36 ± 2.60	83.08	92.53
$+\mathcal{L}_{saac}$	84.28 ± 11.56	96.13 ± 1.35	**81.95 ± 12.60**	88.18 ± 3.50	84.59 ± 8.11	92.95 ± 3.30	86.49 ± 9.69	93.27 ± 3.40	84.33	92.63
$+\mathcal{L}_{dsct}+\mathcal{L}_{saac}$ (CDDSA⋄)	85.53 ± 11.37	96.74 ± 1.70	76.39 ± 17.45	88.25 ± 6.91	84.60 ± 7.76	92.34 ± 4.08	86.56 ± 9.75	92.63 ± 3.39	83.27	92.49
$+\mathcal{L}_{dsct}+\mathcal{L}_{saac}$ (CDDSA)	**85.75 ± 12.31**	**96.79 ± 1.53**	81.04 ± 13.63	89.71 ± 3.60	**86.94 ± 7.94**	93.25 ± 3.55	86.86 ± 8.97	**94.44 ± 3.96**	**85.15**	**93.55**

Table 4
Comparison of ASSD (pixel) by different DG methods on the multi-site fundus image dataset.

Methods	Domain 1		Domain 2		Domain 3		Domain 4		Avg	
	Cup	Disc	Cup	Disc	Cup	Disc	Cup	Disc	Cup	Disc
Lower bound (Inter-domain)	22.35 ± 9.74	6.47 ± 3.80	15.77 ± 20.21	18.25 ± 19.60	12.30 ± 5.82	12.33 ± 5.03	7.45 ± 4.60	9.27 ± 2.62	14.47	11.58
Upper bound (Intra-domain)	16.04 ± 6.65	7.84 ± 3.87	13.10 ± 7.68	8.55 ± 5.80	8.41 ± 5.02	6.32 ± 4.02	6.07 ± 3.41	5.46 ± 2.48	10.91	7.04
BigAug (Zhang et al., 2020)	17.91 ± 10.11	8.67 ± 4.08	22.33 ± 15.26	19.77 ± 6.69	13.51 ± 7.67	14.46 ± 4.96	8.90 ± 5.02	8.77 ± 6.63	15.66	12.92
DoFE (Wang et al., 2020)	17.16 ± 9.40	7.62 ± 2.38	15.28 ± 12.94	14.52 ± 5.36	10.73 ± 6.22	10.11 ± 5.11	**7.18 ± 3.23**	7.60 ± 3.64	12.59	9.96
FedDG (Liu et al., 2021)	18.97 ± 12.82	7.83 ± 3.11	15.34 ± 9.33	13.74 ± 6.79	12.21 ± 5.57	9.71 ± 5.63	9.21 ± 6.62	8.15 ± 5.89	13.93	9.86
DCA-Net (Gu et al., 2021)	17.19 ± 7.64	9.32 ± 4.70	12.39 ± 12.32	10.46 ± 3.23	11.28 ± 5.21	11.32 ± 5.54	7.37 ± 6.51	7.22 ± 4.75	12.06	9.58
Baseline	17.33 ± 8.58	8.21 ± 4.61	13.10 ± 6.66	**11.79 ± 3.90**	11.03 ± 5.22	**9.31 ± 4.23**	8.04 ± 6.42	7.43 ± 4.89	12.38	9.18
$+\mathcal{L}_{dsct}$	18.21 ± 8.05	7.52 ± 5.85	13.33 ± 13.43	14.10 ± 13.50	10.09 ± 5.42	9.98 ± 3.11	7.30 ± 3.88	7.03 ± 2.57	12.23	9.66
$+\mathcal{L}_{saac}$	15.77 ± 6.93	7.14 ± 2.59	**11.04 ± 5.91**	12.97 ± 3.94	10.58 ± 5.21	9.60 ± 3.71	7.22 ± 5.31	7.51 ± 4.10	11.15	9.31
$+\mathcal{L}_{dsct}+\mathcal{L}_{saac}$ (CDDSA⋄)	14.85 ± 6.87	6.78 ± 3.47	15.35 ± 12.98	14.61 ± 12.51	10.72 ± 5.25	9.92 ± 4.07	7.35 ± 4.44	7.75 ± 3.77	12.07	9.77
$+\mathcal{L}_{dsct}+\mathcal{L}_{saac}$ (CDDSA)	**14.65 ± 8.39**	**6.54 ± 3.74**	12.91 ± 10.79	13.06 ± 8.60	**9.38 ± 5.40**	9.32 ± 4.11	7.28 ± 5.85	6.87 ± 5.03	**11.06**	**8.95**

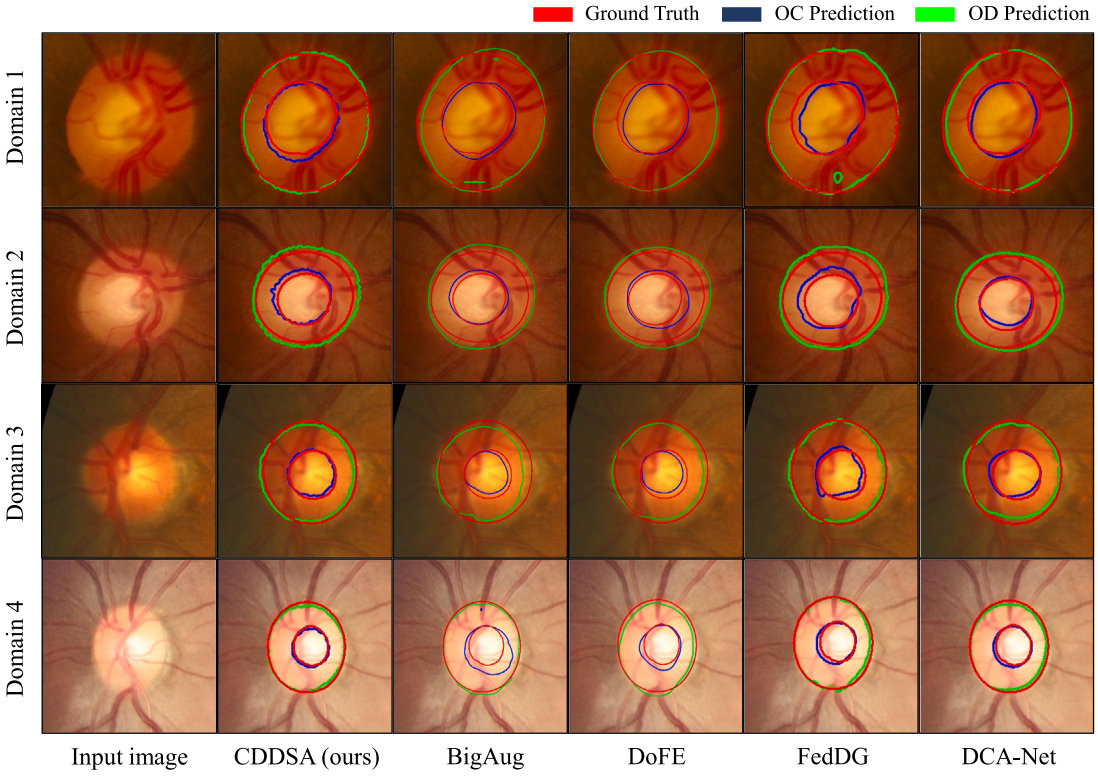

Fig. 4. Visual comparison between our proposed CDDSA and BigAug (Zhang et al., 2020), DoFE (Wang et al., 2020), FedDG (Liu et al., 2021) and DCA-Net (Gu et al., 2021) on multi-domain fundus image segmentation.

DCA-Net (Gu et al., 2021) based on domain composition and attention, and FedDG (Liu et al., 2021) that is a federated learning-based domain generalization method. Here, it just acts as a reference method since the different training strategy from other methods, but some existing works used it for comparison (Hu et al., 2022a; Zhou et al., 2022b,c).

Tables 3 and 4 show the quantitative evaluation results of OC/OD segmentation in terms of Dice and ASSD, respectively. Intra-domain achieved the highest performance among the compared methods, with an average Dice of 85.33% and 95.55% for the OC and OD across the four domains. In contrast, the average Dice achieved by Inter-domain was only 79.61% and 90.32% in OC and OD segmentation, respectively, showing the performance gap caused by domain shift. BigAug obtained a slight improvement from Inter-domain, suggesting that aimlessly conducting data augmentation in the image domain has limited performance. DoFE and FedDG performed better than BigAug. Note that the results of DoFE in our experiment were slightly worse than those in the original paper, as the original implementation used pre-trained weights while the implementation in this work did not. Among the compared existing methods, DCA-Net achieved the highest performance, with an average Dice of 83.60% and 92.64% for OC and OD, respectively. In contrast, our proposed CDDSA outperformed the existing methods, with an average Dice of 85.15% and 93.55% for OC and OD, respectively. The average ASSD obtained by our method was 11.06 and 8.95 pixels for OC and OD, respectively, which also outperformed the compared methods, as shown in Table 4. Fig. 4 shows a visual comparison between our proposed CDDSA and BigAug, DoFE, FedDG and DCA-Net for images from the four testing domains, respectively. It shows that the segmentation results obtained by our proposed CDDSA had boundaries that are closer to the ground truth, while the other DG methods have more over- and under-segmented regions than ours.

4.2.2. Ablation studies

Effectiveness of Domain Style Contrastive Learning and Style Augmentation: We conducted ablation studies to evaluate the effectiveness of the components of our CDDSA framework, where the baseline was only training E_{ana}, E_{sty}, D_{rec} and S with basic loss functions of $\mathcal{L}_{seg}$, $\mathcal{L}_{kl}$ and $\mathcal{L}_{rec}$, following SDNet (Chartsias et al., 2019). We use $+\mathcal{L}_{dsct}$ and $+\mathcal{L}_{saac}$ to denote adding the domain style contrastive learning and domain style augmentation with anatomical consistency to the baseline, respectively. $+\mathcal{L}_{dsct}$ $+\mathcal{L}_{saac}$ means our proposed CDDSA.

Quantitative evaluation results in terms of Dice and ASSD of these variants are shown in the last section of Tables 3 and 4, respectively. The baseline's average Dice across OC and OD was 87.42%, and combining the baseline with $\mathcal{L}_{dsct}$ improved it to 88.81%, indicating that encouraging the network to obtain more discriminative style codes leads to better generalization performance. Baseline + $\mathcal{L}_{saac}$ also improved the two classes' average Dice to 88.48%, and our method using $\mathcal{L}_{dsct}$ and $\mathcal{L}_{saac}$ further improved the average Dice to 89.35% (85.15% for OC and 93.55% for OD), showing the extra improvement brought by the proposed style augmentation.

To show the effectiveness of our proposed domain style contrastive learning, we visualized the T-SNE maps of style codes in different domains with and without contrastive learning, respectively. Here we used domain1 as the unseen target domain, and the visual comparison of source domains is shown in Fig. 6. It shows that when the domain style contrastive learning is not used, the style codes are not well clustered with large inter-domain overlaps and intra-domain divergence. In contrast, after introducing domain style contrastive learning, the style codes have a higher intra-domain compactness and inter-domain distance, making them more distinctive from each other.

To additionally evaluate the effectiveness of our proposed random linear combination for generating new style code during style augmentation, we compared it with an alternative method that randomly samples the style code from a Gaussian distribution, which is denoted as CDDSA$\diamond$ in Tables 3 and 4. The results showed CDDSA$\diamond$ performed slightly worse than CDDSA, but still outperformed most existing DG methods, proving that our proposed random linear combination was better for style augmentation than randomly sampling the style code from a Gaussian distribution.

Reconstruction and Style Augmentation Quality: Since the decomposed anatomical representation f_a serves as the input of the segmentor S and the reconstruction decoder D_{rec}, the quality of the anatomical representation has an impact on the performance of S and D_{rec}. To explore the influence of different formats of the anatomical representation on reconstruction and segmentation quality, we compared four activation functions at the end of E_{ana}: (1) the Gumbel softmax (Jang et al., 2017) returning discrete one-hot values, which is referred to as Gumbel-H; (2) the Gumbel softmax returning continuous soft values, which is referred to as Gumbel-S; (3) Softmax and (4) Tanh. Quantitative comparison of these actuation functions in the fundus image segmentation task is shown in Table 5. We found that Gumbel-S obtained a better segmentation performance than Gumbel-H (83.56% and 92.31% vs. 83.18% and 93.20% of average Dice score). Softmax and Tanh further improved the model's performance. Notably, Tanh achieved the highest average Dice score of 85.15% in OC and 93.55% in OD and the lowest average ASSD (11.06 pixels in OC and 8.95 pixels in OD) among the compared activation functions. The results show that using continuous soft values for anatomical representations led to better segmentation performance, as soft representations are more informative compared with binary representations (Chartsias et al., 2020).

To further investigate how the activation function used by E_{ana} affects the reconstructed images and style augmentation, we compared the original training image with the image reconstructed from disentangled f_a and f_s and the style-augmented image in Fig. 5, where we show the differences between Gumbel-H, Gumbel-S and Tanh. First, for image reconstruction (even rows), it can be observed that using Gumbel-H only reconstructed coarse-grained images and only roughly retrained the overall content without detailed structures. Gumbel-S achieved a better quality with more details than Gumbel-H. However, the reconstructed images have some noticeable artefacts compared with the original images. In contrast, Tanh obtained a much higher quality than Gumbel-H, and the reconstructed images were closer to the original inputs in terms of both anatomical structures and styles. Second, for style augmentation (odd rows), Gumbel-H cannot keep the same anatomical structure after changing the style, and led to unrealistic images in the augmented domain, especially in cases 5, 7, 8 and 10 in the third row of Fig. 5. Gumbel-S has a better ability to retain the anatomical structures, but the augmented images have a lot of artefacts with unrealistic appearances. In contrast, Tanh achieved very high quality in the style-augmented images with realistic appearances. They have quite different styles with shared anatomical structures compared with the original images, as shown in the last row of Fig. 5. The results show the advantage of our proposed style augmentation strategy, which can successfully generate new samples in an unknown domain with anatomical structures unchanged, which is beneficial for enhancing the model's generalizability.

Effectiveness of Disentanglement: To assess the successful disentanglement of anatomical representations, we performed an exchange of the decomposed style codes between images from different domains. For an image from a certain domain, we reconstruct an output based on its anatomical representation and style codes from different domains, respectively. The results are shown in Fig. 7. It can be observed that the disentanglement network effectively reconstructs the original image while preserving an adequate amount of anatomical and style information. Upon exchanging the style codes, the reconstructed images exhibit a transfer of styles to the corresponding domain, as illustrated in the non-diagonal samples of Fig. 7. Importantly, these style-exchanged images retain all intrinsic anatomical information, indicating the successful disentanglement of anatomical representations through our proposed disentanglement method.

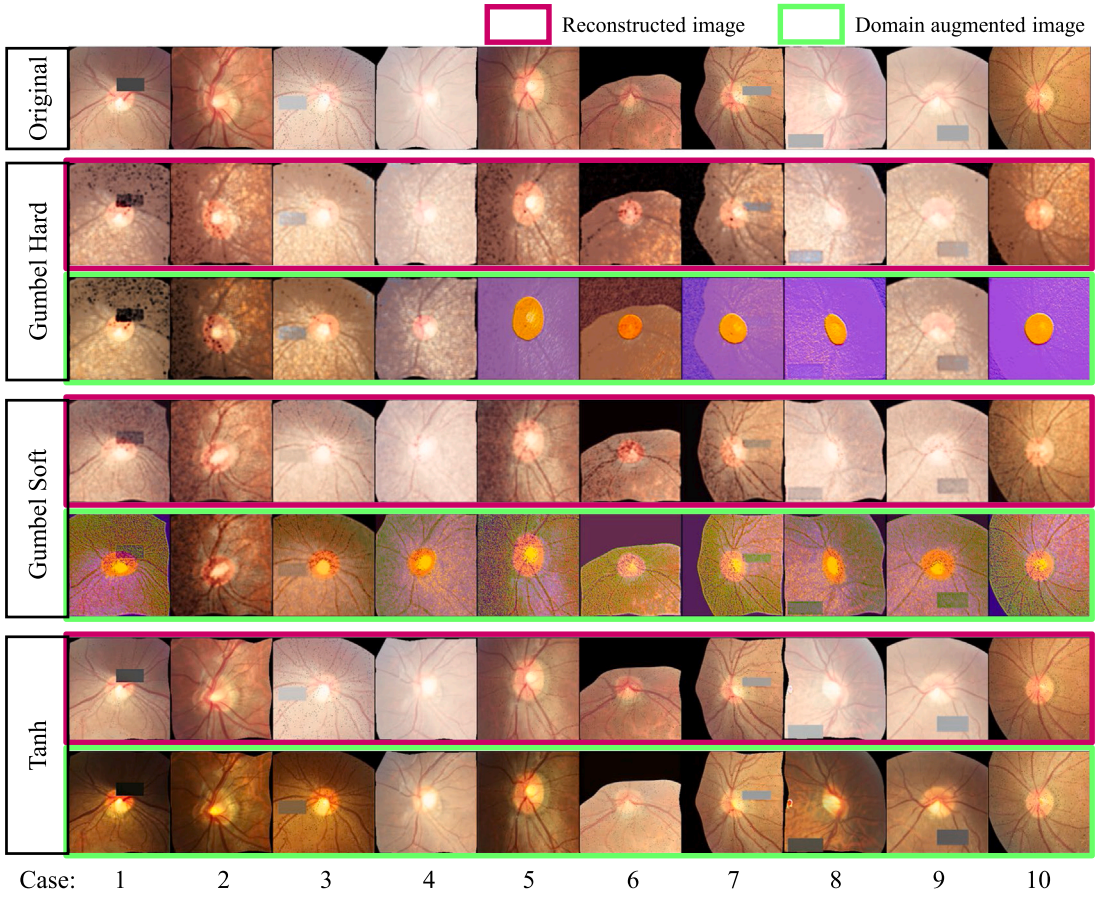

Fig. 5. Visual comparison of reconstructed and augmented fundus images with different activation functions at the end of the anatomy Encoder. The original images are from domain 4. For each method, the first row (red rectangles) shows images reconstructed from the disentangled anatomical representation and style code, and the second row (green rectangles) shows style-augmented images that are generated based on the anatomical representation from the original images and changed style codes.

Table 5
Comparison between different activation functions used by the output of E_{ana} for multi-domain OC/OD segmentation. Gumbel-H and Gumbel-S are two variants of Gumbel softmax that return discrete one-hot values and soft continuous values, respectively.

Metric	Activation	Domain 1		Domain 2		Domain 3		Domain 4		Avg	
		Cup	Disc	Cup	Disc	Cup	Disc	Cup	Disc	Cup	Disc
Dice	Gumbel-H	82.29 ± 11.97	96.46 ± 2.11	78.67 ± 18.79	86.21 ± 5.06	85.40 ± 8.14	**93.36 ± 3.19**	**87.89 ± 7.67**	93.22 ± 3.16	83.56	92.31
	Gumbel-S	82.88 ± 11.28	96.71 ± 1.65	80.22 ± 13.76	89.11 ± 3.52	83.20 ± 8.90	92.59 ± 4.91	86.41 ± 9.88	94.39 ± 3.53	83.18	93.20
	Softmax	85.22 ± 10.46	**96.94 ± 1.26**	80.72 ± 16.09	88.42 ± 12.16	85.22 ± 7.43	93.13 ± 3.75	85.34 ± 10.07	93.23 ± 3.78	84.13	92.93
	Tanh	**85.75 ± 12.31**	96.79 ± 1.53	**81.04 ± 13.63**	**89.71 ± 3.60**	**86.94 ± 7.94**	93.25 ± 3.55	86.86 ± 8.97	**94.44 ± 3.96**	**85.15**	**93.55**
ASSD	Gumbel-H	18.31 ± 8.49	6.65 ± 4.11	14.60 ± 17.21	18.13 ± 16.60	9.92 ± 5.45	**9.10 ± 3.66**	**6.67 ± 3.96**	6.97 ± 3.02	12.38	10.21
	Gumbel-S	17.22 ± 6.91	**6.33 ± 2.83**	**12.30 ± 6.02**	**12.98 ± 5.03**	10.94 ± 5.57	9.78 ± 5.85	7.52 ± 5.42	6.89 ± 4.77	12.00	9.00
	Softmax	**14.40 ± 6.36**	6.67 ± 2.53	12.61 ± 12.71	13.28 ± 12.93	10.01 ± 5.14	9.90 ± 5.09	8.28 ± 5.90	7.67 ± 4.73	11.33	9.38
	Tanh	14.65 ± 8.39	6.54 ± 3.74	12.91 ± 10.79	13.06 ± 8.60	**9.38 ± 5.40**	9.32 ± 4.11	7.28 ± 5.85	**6.87 ± 5.03**	**11.06**	**8.95**

Table 6
Quantitative comparison of different DG methods on the multi-domain NPC-MRI image dataset for GTVnx segmentation.

Methods	Domain 1		Domain 2		Domain 3		Domain 4		Avg	
	Dice (%)	ASSD (mm)	Dice (%)	ASSD (mm)	Dice (%)	ASSD (mm)	Dice (%)	ASSD (mm)	Dice	ASSD
Lower bound (Inter-domain)	65.44 ± 13.35	2.37 ± 1.52	76.65 ± 7.75	1.70 ± 1.06	81.17 ± 7.01	1.61 ± 0.77	61.62 ± 13.66	3.38 ± 1.47	71.22	2.27
Upper bound (Intra-domain)	79.19 ± 6.35	1.13 ± 0.54	82.89 ± 7.54	1.58 ± 1.75	86.30 ± 3.25	1.26 ± 0.48	79.46 ± 7.40	2.09 ± 1.48	81.96	1.52
BigAug (Zhang et al., 2020)	75.63 ± 5.97	1.67 ± 1.01	78.51 ± 7.89	1.65 ± 0.99	82.30 ± 5.05	1.82 ± 0.66	63.88 ± 12.32	4.05 ± 1.71	75.08	2.30
DoFE (Wang et al., 2020)	**78.44 ± 6.97**	**1.27 ± 0.99**	75.00 ± 4.95	1.80 ± 0.85	79.66 ± 5.77	1.66 ± 0.68	64.71 ± 15.06	2.57 ± 1.91	74.45	1.83
FedDG (Liu et al., 2021)	65.07 ± 11.59	2.58 ± 1.63	78.90 ± 6.11	1.67 ± 1.09	81.57 ± 6.05	1.79 ± 0.79	67.58 ± 12.55	6.67 ± 3.61	73.28	3.18
DCA-Net (Gu et al., 2021)	77.27 ± 6.66	1.27 ± 0.99	77.14 ± 7.53	1.80 ± 0.85	81.63 ± 6.20	1.66 ± 0.68	69.32 ± 10.08	2.57 ± 1.91	76.34	1.83
Baseline	76.63 ± 5.74	1.53 ± 1.23	77.93 ± 5.99	**1.41 ± 0.72**	82.77 ± 4.65	1.55 ± 0.55	62.71 ± 11.63	3.07 ± 1.91	75.01	1.89
$+\mathcal{L}_{dsct}$	77.02 ± 6.09	1.56 ± 1.09	77.11 ± 7.15	1.71 ± 0.89	83.02 ± 3.62	1.45 ± 0.41	63.88 ± 12.74	2.81 ± 1.66	75.26	1.88
$+\mathcal{L}_{saac}$	77.74 ± 5.67	1.35 ± 0.82	76.33 ± 8.37	1.66 ± 0.83	83.23 ± 5.38	**1.44 ± 0.56**	68.43 ± 12.92	3.45 ± 1.99	76.43	1.98
$+\mathcal{L}_{dsct}+\mathcal{L}_{saac}$ (CDDSA$\diamond$)	77.77 ± 6.65	1.36 ± 0.97	78.00 ± 7.74	1.52 ± 0.86	83.10 ± 5.18	1.60 ± 0.63	66.39 ± 10.88	3.50 ± 1.62	76.32	2.00
$+\mathcal{L}_{dsct}+\mathcal{L}_{saac}$ (CDDSA)	78.34 ± 5.14	1.37 ± 0.82	**79.16 ± 6.68**	1.61 ± 1.19	**83.53 ± 4.55**	1.48 ± 0.54	**69.53 ± 10.28**	**2.46 ± 1.50**	**77.64**	**1.73**

4.3. NPC GTVnx image segmentation

4.3.1. Comparison with state-of-the-art DG methods

For the multi-domain GTVnx segmentation task, we employed the same set of methods as in Section 4.2.1 for comparison, and the quantitative evaluation results are shown in Table 6. First, Intra-domain (upper bound) achieved the highest performance with average Dice of 81.96% and average ASSD of 1.52 mm across the four domains. In contrast, Inter-domain (lower bound) only obtained an average Dice of 71.22% and ASSD of 2.27 mm. The performance gap between them was over 10% in average Dice, indicating the large shift among the different domains. All four existing DG methods achieved great improvements

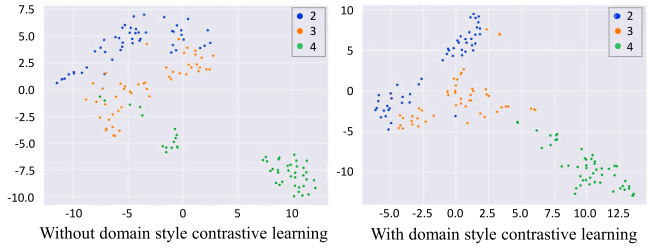

Fig. 6. Visual comparison of T-SNE on extracted style codes before and after using domain style contrastive learning. The style codes are from domain 2, 3 and 4 of the fundus image dataset.

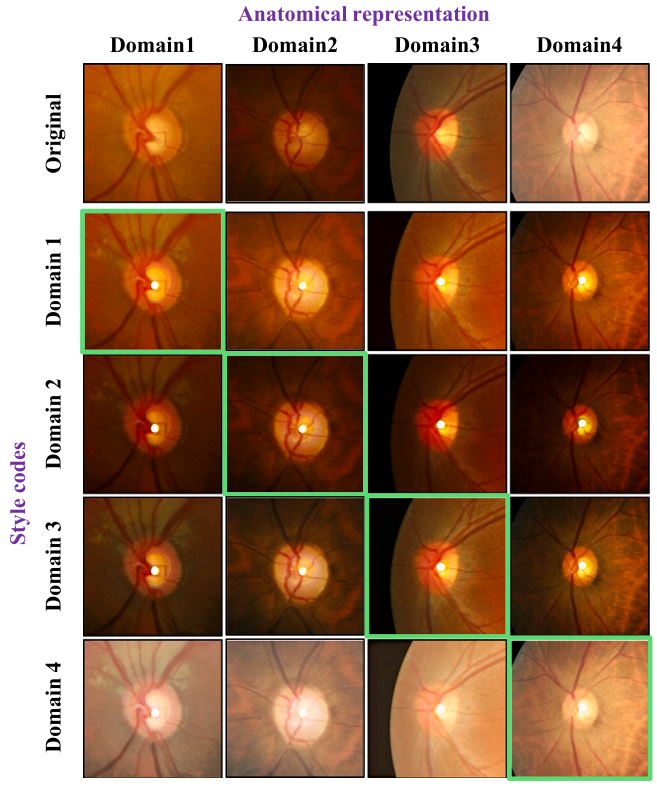

Fig. 7. Visual comparison between reconstructed images based on styles decomposed from different domains. Green boxes correspond to the reconstruction of the original images in each domain.

compared with the Inter-domain that does not consider the differences across domains. DoFE (Wang et al., 2020) and FedDG (Liu et al., 2021) had a similar segmentation performance, with average Dice scores of 74.45% and 74.48%, respectively. BigAug (Zhang et al., 2020) achieved an average Dice of 75.08%, indicating that some augmentation strategies are beneficial for GTVnx segmentation in cross-modality MRI images. DCA-Net (Gu et al., 2021) obtained an average Dice of 76.34%, which outperformed the other three existing DG methods. In contrast, our proposed CDDSA based on domain-invariant feature learning obtained higher generalizability, achieving an average Dice of 77.64% and ASSD of 1.73 mm, which outperformed the state-of-the-art DG methods.

Fig. 8 provides a visual comparison between our proposed CDDSA and the four state-of-the-art DG methods on the multi-domain NPC GTVnx segmentation dataset. Fig. 8(a) shows that the 2D segmentation boundaries of our CDDSA are closer to the ground truth than those of the other methods. The 3D visualization in Fig. 8(b) shows that our CDDSA achieved high-quality segmentation results, while the other DG methods have more noises in the results.

4.3.2. Ablation studies

Effectiveness of Domain Style Contrastive Learning and Style Augmentation: Similar with multi-site fundus image segmentation. We also proved the effectiveness of our proposed domain style contrastive learning and style augmentation strategy in multi-site NPC GTVnx segmentation. Quantitative results are shown in Table 6. The baseline, i.e., re-implementation of SDNet (Chartsias et al., 2019) based on our network structures, obtained an average Dice of 75.01%, and combining it with our domain style contrastive learning $\mathcal{L}_{dsct}$ improved it to 75.26%. Combining it with our domain augmentation method $\mathcal{L}_{saac}$ achieved an average Dice of 76.43%. In contrast, our proposed method that uses $\mathcal{L}_{dsct}$ and $\mathcal{L}_{saac}$ simultaneously improved the average Dice to 77.64%, which is the highest among the compared variants and significantly better than the baseline (p-value < 0.05). Table 6 also shows that CDDSA performed better than CDDSA$\diamond$ (77.64% vs. 76.32%) in terms of Dice, indicating that our style augmentation based on random linear combination of the style codes was better than directly sampling style codes from a Gaussian distribution for style augmentation.

Reconstruction and Style Augmentation Qualities: Similar to Section 4.2.2, we compared four different activation functions at the end of E_{ana} to represent f_a on the NPC-MRI dataset. The corresponding NPC-MRI GTVnx segmentation results are shown in Table 7. It can be observed that Gumbel-S had a higher performance than Gumbel-H (75.88% vs. 73.96% in terms of average Dice). Using Tanh further improved the average Dice to 77.64%, which was significantly better than the other activations.

Fig. 9 shows a visual comparison of these activation functions in reconstructing the original image after disentanglement. It can be observed that when Gumbel-H is used, the reconstructed images have a large difference from the original images. Gumbel-S has a lower reconstruction error than Gumbel-H. However, it is inferior to our method using Tanh, showing that Tanh is more suitable to obtaining anatomical representation in disentanglement for high-fidelity reconstruction.

In addition, Fig. 10 shows a visual comparison of style-augmented images when different activation functions are used at the end of E_{ana}. We found that all the methods can generate new-style images based on the augmented domain style code $\tilde{f}_s$ and the anatomical representation f_a of the input. However, Gumbel-H and Gumbel-S led to obvious artefacts in the augmented images. In contrast, our method can change the style of an input image while better retaining the anatomical structures.

5. Discussion

Disentanglement network is an intuitive strategy to address domain generalization, where enabling the neural networks to extract domain-invariant features can effectively improve their robustness when applied to different domains. Existing disentanglement networks are commonly based on GAN (Pei et al., 2021; Xie et al., 2022). However, they commonly require domain-specific content and style encoders and are not suitable for the unseen target domains where the data are not available in advance. Hence, GAN-based disentanglement may not be adaptable for domain generalization. In this work, we adopt the disentanglement with sharing the encoders across domains, which is especially adaptable for the scenario of multi-modality segmentation and unpredictable target domains. Additionally, we discovered that using soft values to represent anatomical features is beneficial for reconstruction. The reason is mainly that soft values are capable of sufficiently restoring the details of anatomical structures.

In our method, the style augmentation technique plays a crucial role in ensuring anatomical consistency and learning domain-invariant anatomical features. We found that random sampling from a Gaussian distribution for style augmentation is not as effective as a random linear combination of disentangled style codes. This is mainly due to the fact that the actual distribution of style codes in the dataset may deviate

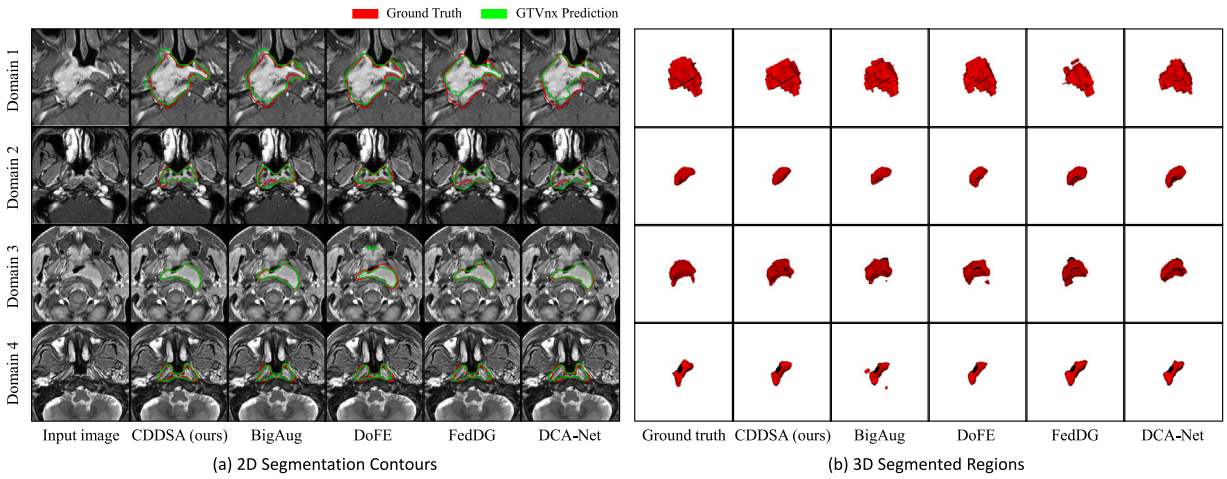

Fig. 8. Visual comparison between different DG methods for multi-domain NPC GTVnx segmentation.

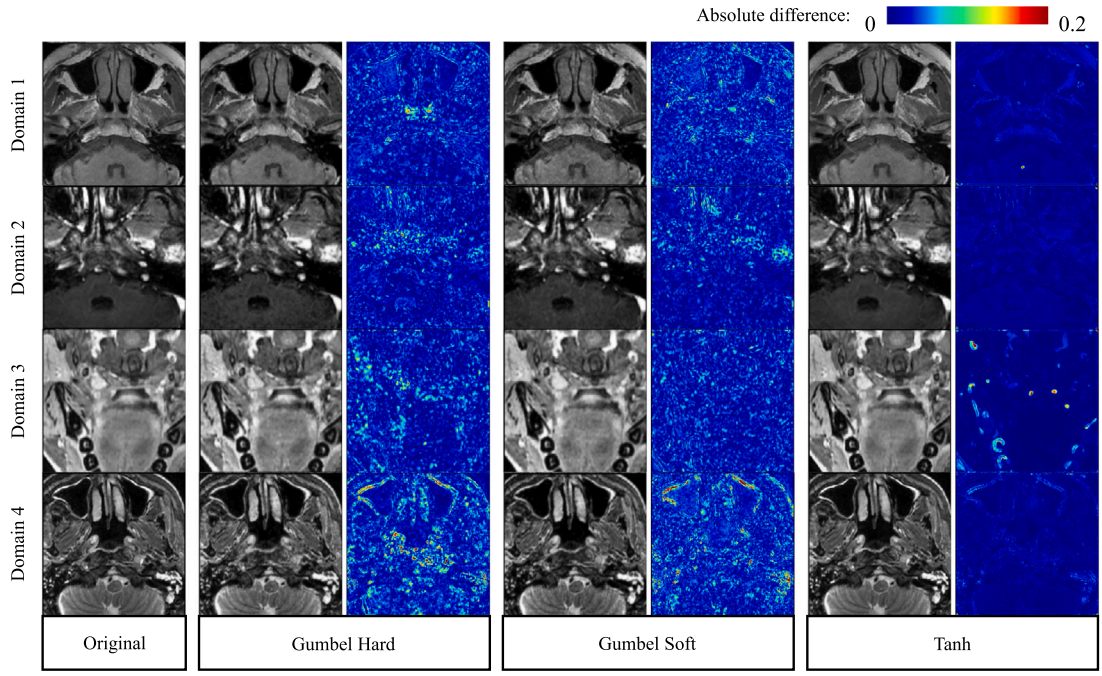

Fig. 9. Comparison of reconstructed images with different activation functions at the end of E_{ana}. For each method, the first column shows the reconstructed images based on the disentangled anatomical representation and style code, and the second column shows the absolute difference between the reconstructed and original images.

Table 7
Comparison between different activation functions used by the output of E_{ana} for multi-domain NPC GTVnx segmentation.

Activation	Domain 1		Domain 2		Domain 3		Domain 4		Avg	
	Dice (%)	ASSD (mm)	Dice (%)	ASSD (mm)	Dice (%)	ASSD (mm)	Dice (%)	ASSD (mm)	Dice (%)	ASSD (mm)
Gumbel-H	75.60 ± 6.20	1.61 ± 1.15	72.69 ± 9.42	1.60 ± 0.76	82.81 ± 7.67	1.62 ± 0.58	64.73 ± 13.53	2.90 ± 1.88	73.96	1.93
Gumbel-S	76.54 ± 6.39	1.48 ± 1.10	78.40 ± 6.48	1.57 ± 0.88	83.23 ± 4.54	**1.47 ± 0.52**	65.24 ± 11.78	2.73 ± 1.43	75.88	1.81
Softmax	77.46 ± 5.40	1.57 ± 1.01	**79.87 ± 4.70**	**1.46 ± 0.74**	82.85 ± 4.89	1.89 ± 0.91	62.59 ± 12.48	3.08 ± 1.59	75.69	2.00
Tanh	**78.34 ± 5.14**	**1.37 ± 0.82**	79.16 ± 6.68	1.61 ± 1.19	**83.53 ± 4.55**	1.48 ± 0.54	**69.53 ± 10.28**	**2.46 ± 1.50**	**77.64**	**1.73**

from an ideal Gaussian distribution. By utilizing a linear combination of existing style codes to generate new ones, our method better adapts to the specific segmentation task at hand, resulting in higher-quality image synthesis during style augmentation.

The introduction of contrastive learning on domain style improves model generalization mainly due to two reasons. First, it minimizes the similarity of style code in different domains, leading to a better disentanglement for separating the domain-specific and domain-invariant features that can improve generalizability. Second, when the decomposed style codes are more discriminative based on contrastive learning, their combination can obtain more diverse new styles, leading to better style augmentation results that help to improve model generalization.

Compared to other state-of-the-art methods, our proposed CDDSA offers the ability to synthesize images with various styles in real time, with only a minimal increase in time consumption. Considering domain 1 as the target domain in the context of fundus segmentation, the lower-bound method, which does not consider domain generalization, requires approximately 80 min for the training phase. When considering enhanced domain generalization, state-of-the-art methods such as BigAug, DoFE, FedDG, and DCA-Net take approximately 156,

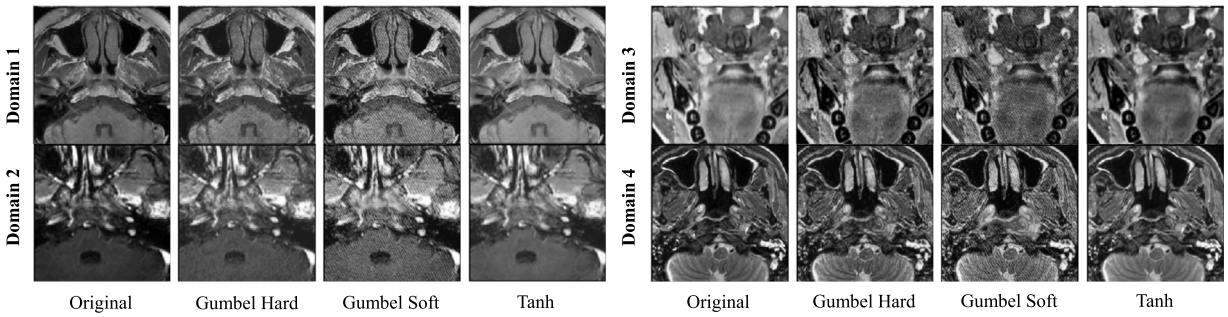

Fig. 10. Visual comparison of style-augmented images with different activation functions at the end of E_{ana}.

118, 960, and 90 min, respectively. In comparison, CDDSA requires approximately 140 min for the training phase, adding only additional 53 min compared to the baseline method SDNet. Therefore, when compared to these state-of-the-art domain generalization methods, our proposed CDDSA exhibits superior model performance with reasonable time consumption.

One limitation of this work is that our proposed CDDSA was only applied to 2D images. In future work, we aim to extend the CDDSA framework to 3D images and validate its effectiveness on a diverse and extensive collection of medical images. Another limitation is that this work requires access to all the source domains at the same time, which may increase the difficulty of data preparation in the training phase. To provide a more flexible domain generalization method, it is promising to consider using the federated learning strategy that does not need to share data across centres.

6. Conclusion

In this paper, we present a Contrastive Domain Disentanglement and Style Augmentation (CDDSA) framework to tackle the domain generalization problem in medical image segmentation. We introduce a GAN-free efficient disentangle method to decompose medical images from multiple domains into a domain-invariant anatomical representation and a domain-specific style code, where a segmentor works on the anatomical representation to achieve generalizability. To improve the disentanglement and segmentation performance, we use a soft representation for the anatomical representation based on Tanh and propose domain style contrastive learning to minimize the similarity of style codes in different domains. Based on the disentanglement, we propose a style augmentation strategy that changes the style of an image with remaining structure information for augmentation, which can further improve the model's generalizability. Quantitative experimental results on a multi-site fundus image dataset and a multi-domain NPC MRI dataset showed that our CDDSA outperformed several state-of-the-art multi-domain generalization methods. In the future, it is of interest to apply our CDDSA framework to other multi-domain medical image analysis tasks.

CRediT authorship contribution statement

Ran Gu: Conceptualization, Methodology, Software, Writing – original draft, Visualization. **Guotai Wang:** Conceptualization, Methodology, Writing – review & editing, Supervision. **Jiangshan Lu:** Software, Visualization. **Jingyang Zhang:** Conceptualization, Methodology, Writing – review & editing. **Wenhui Lei:** Conceptualization, Methodology, Writing – review & editing. **Yinan Chen:** Resources. **Wenjun Liao:** Data Curation, Resources. **Shichuan Zhang:** Data Curation, Resources. **Kang Li:** Resources. **Dimitris N. Metaxas:** Resources. **Shaoting Zhang:** Methodology, Resources, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yinan Chen is employed by SenseTime at the time of submission.

Data availability

We have shared the link to my code in the manuscript.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62271115), National Key Research and Development Program of China (2020YFB1711500), the 1·3·5 project for disciplines of excellence, West China Hospital, Sichuan University (ZYYC21004) and Radiation Oncology Key Laboratory of Sichuan Province Open Fund (2022ROKF04).

References

Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 35 (8), 1798–1828.

Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E., 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. Adv. Neural Inf. Process. Syst. 33.

Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D.E., Dharmakumar, R., Tsaftaris, S.A., 2019. Disentangled representation learning in cardiac image analysis. Med. Image Anal. 58, 101535.

Chartsias, A., Papanastasiou, G., Wang, C., Semple, S., Newby, D.E., Dharmakumar, R., Tsaftaris, S.A., 2020. Disentangle, align and fuse for multimodal and semi-supervised image segmentation. IEEE Trans. Med. Imaging 40 (3), 781–792.

Chen, C., Dou, Q., Chen, H., Heng, P.-A., 2018. Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest X-Ray segmentation. In: Shi, Y., Suk, H.-I., Liu, M. (Eds.), Machine Learning in Medical Imaging. Springer International Publishing, Cham, pp. 143–151.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. PMLR, pp. 1597–1607.

Dou, Q., Coelho de Castro, D., Kamnitsas, K., Glocker, B., 2019. Domain generalization via model-agnostic learning of semantic features. Adv. Neural Inf. Process. Syst. 32.

Duan, Q., Wang, G., Wang, R., Fu, C., Li, X., Wang, N., Huang, Y., Huang, X., Song, T., Zhao, L., Liu, X., Xia, Q., Hu, Z., Chen, Y., Zhang, S., 2020. SenseCare: A research platform for medical image informatics and interactive 3D visualization. arXiv:2004.07031.

Fick, R.H., Moshayedi, A., Roy, G., Dedieu, J., Petit, S., Hadj, S.B., 2021. Domain-specific cycle-GAN augmentation improves domain generalizability for mitosis detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 40–47.

Fumero, F., Alayón, S., Sanchez, J.L., Sigut, J., Gonzalez-Hernandez, M., 2011. RIM-ONE: An open retinal image database for optic nerve evaluation. In: 2011 24th International Symposium on Computer-Based Medical Systems (CBMS). IEEE, pp. 1–6.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. J. Mach. Learn. Res. 17 (1), 189–209.

Gatys, L.A., Ecker, A.S., Bethge, M., 2016. Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2414–2423.

Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S., 2021. CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Trans. Med. Imaging 40 (2), 699–711.

Gu, R., Zhang, J., Huang, R., Lei, W., Wang, G., Zhang, S., 2021. Domain composition and attention for unseen-domain generalizable medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 241–250.

Gu, R., Zhang, J., Wang, G., Lei, W., Song, T., Zhang, X., Li, K., Zhang, S., 2022. Contrastive semi-supervised learning for domain adaptive segmentation across similar anatomical structures. IEEE Trans. Med. Imaging 1.

Guan, H., Liu, M., 2021. Domain adaptation for medical image analysis: a survey. IEEE Trans. Biomed. Eng. 69 (3), 1173–1185.

Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2. IEEE, pp. 1735–1742.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738.

Hu, S., Liao, Z., Zhang, J., Xia, Y., 2022a. Domain and content adaptive convolution based multi-source domain generalization for medical image segmentation. IEEE Trans. Med. Imaging 42 (1), 233–244.

Hu, S., Liao, Z., Zhang, J., Xia, Y., 2022b. Domain and content adaptive convolution for domain generalization in medical image segmentation. IEEE Trans. Med. Imaging 1.

Huang, X., Liu, M.-Y., Belongie, S., Kautz, J., 2018. Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 172–189.

Jang, E., Gu, S., Poole, B., 2017. Categorical reparameterization with gumbel-softmax. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al., 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 597–609.

Kang, G., Jiang, L., Yang, Y., Hauptmann, A.G., 2019. Contrastive adaptation network for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4893–4902.

Kingma, D.P., Welling, M., 2014. Auto-encoding variational bayes. In: International Conference on Learning Representations.

Lei, W., Xu, W., Gu, R., Fu, H., Zhang, S., Zhang, S., Wang, G., 2021. Contrastive learning of relative position regression for one-shot object localization in 3D medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 155–165.

Li, Y., Gong, M., Tian, X., Liu, T., Tao, D., 2018c. Domain generalization via conditional invariant representations. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32.

Li, C., Lin, X., Mao, Y., Lin, W., Qi, Q., Ding, X., Huang, Y., Liang, D., Yu, Y., 2022. Domain generalization on medical imaging classification using episodic training with task augmentation. Comput. Biol. Med. 141, 105144.

Li, H., Pan, S.J., Wang, S., Kot, A.C., 2018b. Domain generalization with adversarial feature learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5400–5409.

Li, D., Yang, Y., Song, Y.-Z., Hospedales, T., 2018a. Learning to generalize: Meta-learning for domain generalization. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol.32.

Liao, W., He, J., Luo, X., Wu, M., Shen, Y., Li, C., Xiao, J., Wang, G., Chen, N., 2022. Automatic delineation of gross tumor volume based on magnetic resonance imaging by performing a novel semi-supervised learning framework in nasopharyngeal carcinoma. Int. J. Radiat. Oncol. Biol. Phys..

Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P.-A., 2021. 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.

Liu, Q., Dou, Q., Heng, P.-A., 2020. Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 475–485.

Meng, Q., Pawlowski, N., Rueckert, D., Kainz, B., 2019. Representation disentanglement for multi-task learning with application to fetal ultrasound. In: Smart Ultrasound Imaging and Perinatal, Preterm and Paediatric Image Analysis. Springer, pp. 47–55.

Muandet, K., Balduzzi, D., Schölkopf, B., 2013. Domain generalization via invariant feature representation. In: International Conference on Machine Learning. PMLR, pp. 10–18.

Ning, M., Bian, C., Wei, D., Yu, S., Yuan, C., Wang, Y., Guo, Y., Ma, K., Zheng, Y., 2021. A new bidirectional unsupervised domain adaptation segmentation framework. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 492–503.

Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Orlando, J.I., Fu, H., Breda, J.B., van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.-A., Kim, J., Lee, J., et al., 2020. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med. Image Anal. 59, 101570.

Pei, C., Wu, F., Huang, L., Zhuang, X., 2021. Disentangle domain features for cross-modality cardiac image segmentation. Med. Image Anal. 71, 102078.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Shen, D., Wu, G., Suk, H.-I., 2017. Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 19, 221–248.

Sivaswamy, J., Krishnadas, S., Chakravarty, A., Joshi, G., Tabish, A.S., et al., 2015. A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. JSM Biomed. Imaging Data Pap. 2 (1), 1004.

Tran, L., Yin, X., Liu, X., 2017. Disentangled representation learning gan for pose-invariant face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1415–1424.

Tzeng, E., Hoffman, J., Saenko, K., Darrell, T., 2017. Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7167–7176.

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., Yu, P., 2022. Generalizing to unseen domains: A survey on domain generalization. IEEE Trans. Knowl. Data Eng..

Wang, S., Yu, L., Li, K., Yang, X., Fu, C.-W., Heng, P.-A., 2020. DoFE: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. IEEE Trans. Med. Imaging.

Wang, G., Zhang, S., Huang, X., Vercauteren, T., Metaxas, D., 2023. Editorial for special issue on explainable and generalizable deep learning methods for medical image computing. Med. Image Anal. 84, 102727.

Wang, X., Zhang, R., Shen, C., Kong, T., Li, L., 2021. Dense contrastive learning for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3024–3033.

Wu, J., Gu, R., Dong, G., Wang, G., Zhang, S., 2022. FPL-UDA: Filtered pseudo label-based unsupervised cross-modality adaptation for vestibular schwannoma segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 1–5.

Xie, X., Chen, J., Li, Y., Shen, L., Ma, K., Zheng, Y., 2020. MI^{2}GAN: Generative adversarial network for medical image domain adaptation using mutual information constraint. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 516–525.

Xie, Q., Li, Y., He, N., Ning, M., Ma, K., Wang, G., Lian, Y., Zheng, Y., 2022. Unsupervised domain adaptation for medical image segmentation by disentanglement learning and self-training. IEEE Trans. Med. Imaging.

Yang, J., Dvornek, N.C., Zhang, F., Chapiro, J., Lin, M., Duncan, J.S., 2019. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 255–263.

You, C., Zhao, R., Staib, L.H., Duncan, J.S., 2022. Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Springer Nature Switzerland, Cham, pp. 639–652.

Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B.J., Roth, H., Myronenko, A., Xu, D., et al., 2020. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE Trans. Med. Imaging 39 (7), 2531–2540.

Zhou, Z., Qi, L., Shi, Y., 2022a. Generalizable medical image segmentation via random amplitude mixup and domain specific image restoration. In: Proceedings of the European Conference on Computer Vision (ECCV).

Zhou, Z., Qi, L., Shi, Y., 2022b. Generalizable medical image segmentation via random amplitude mixup and domain-specific image restoration. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI. Springer, pp. 420–436.

Zhou, Z., Qi, L., Yang, X., Ni, D., Shi, Y., 2022c. Generalizable cross-modality medical image segmentation via style augmentation and dual normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20856–20865.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2223–2232.