Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Research paper

Segmentation-synthesis co-training for semi-supervised domain generalizable medical image segmentation

Zhiqiang Shen [a,b], Qingshan Hou [a,b], Peng Cao [a,b,c,*], Jinzhu Yang [a,b,c], Huazhu Fu [d], Osmar R. Zaiane [e], Zhaolin Chen [f,g]

[a] School of Computer Science and Engineering, Northeastern University, Shenyang, 110819, China
[b] Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Shenyang, 110819, China
[c] National Frontiers Science Center for Industrial Intelligence and Systems Optimization, Shenyang, 110819, China
[d] Institute of High-Performance Computing, Agency for Science, Technology, and Research (A*STAR), Singapore
[e] Alberta Machine Intelligence Institute, University of Alberta, Edmonton, Alberta, Canada
[f] Department of Data Science & AI, Faculty of Information Technology, Monash University, Melbourne, Victoria, 3800, Australia
[g] Monash Biomedical Imaging, Monash University, Melbourne, Victoria, 3800, Australia

ARTICLE INFO

Keywords:
Medical image segmentation
Domain generalization
Semi-supervised learning
Consistency regularization

ABSTRACT

Semi-supervised domain generalization (SSDG) faces two fundamental challenges that hinder model generalizability: label scarcity and domain shifts. Recent studies tackle this challenging task by building on a scheme that integrates the strong–weak pseudo supervision paradigm with specific data augmentation strategies. However, semantic inconsistencies between style-augmented unlabeled images and their pseudo labels limit the effectiveness of this scheme and impair the generalizability of trained models. One critical question arises: *How to ensure semantic consistency and style diversity of unlabeled-image and pseudo-label pairs for training a well-generalized model?* To this end, we introduce **ReMatch**, a segmentation-synthesis co-training framework for semi-supervised domain generalization in medical image segmentation. The core of ReMatch lies in the **SynTS** algorithm that **Syn**thesizes unlabeled images with both high semantic consistency and style diversity by leveraging **T**exture and **S**hape features derived from the segmentation process. Extensive experiments on single-source single-target and single-source multi-target cross-domain settings with various image modalities demonstrate that ReMatch offers an effective solution for SSDG, achieving compelling performance. For example, compared with the state-of-the-art based on the aforementioned scheme, ReMatch achieves average improvements of 2.31% and 1.68% in Dice similarity coefficients under the two cross-domain settings with 10% labeled data, respectively. Code is available at https://github.com/Senyh/ReMatch.

1. Introduction

Label scarcity and domain shifts hinder efforts to improve generalization performance in developing deep learning-based medical image segmentation models for clinical applications [1,2]. **Label scarcity** refers to the limited availability of labeled data due to the time-consuming and expertise-intensive annotation process, whereas **domain shifts** in medical imaging primarily manifest as pixel-intensity disparities due to variations in acquisition systems.

Semi-supervised domain generalization (SSDG) aims to address both issues simultaneously [3,4,6–8]. In this setting, models are trained on partially labeled source domain data and evaluated on unseen target domains. A mainstream scheme for SSDG involves a strong–weak pseudo supervision paradigm [9] to enforce consistency regularization using unlabeled images and pseudo labels, in combination with data augmentation strategies that act as strong perturbations to manipulate image style [3,4,6–8]. Although this scheme enhances *style diversity* in the source domain, it fails to ensure *semantic consistency* between augmented images and pseudo labels, resulting in the model overfitting to inconsistent data with low-quality pseudo labels and impairing model generalization. To investigate the impact of such inconsistencies, we conducted a pilot study on a cross-domain cardiac segmentation task with 10% labeled data. As shown in Fig. 1, while the strong–weak pseudo supervision-based approaches, including FACT-Fix [3], MixStyle-Fix [4], and ConStyX [5] achieve lower training losses, semantic inconsistencies degrade cross-domain segmentation performance. An intuitive solution to this issue is to generate unlabeled-image and pseudo-label pairs with both high semantic consistency and style diversity. However, data augmentation methods, such as

* Corresponding author at: School of Computer Science and Engineering, Northeastern University, Shenyang, 110819, China.
E-mail address: caopeng@cse.neu.edu.cn (P. Cao).

https://doi.org/10.1016/j.artmed.2026.103512
Received 23 March 2026; Received in revised form 24 July 2026; Accepted 12 August 2026

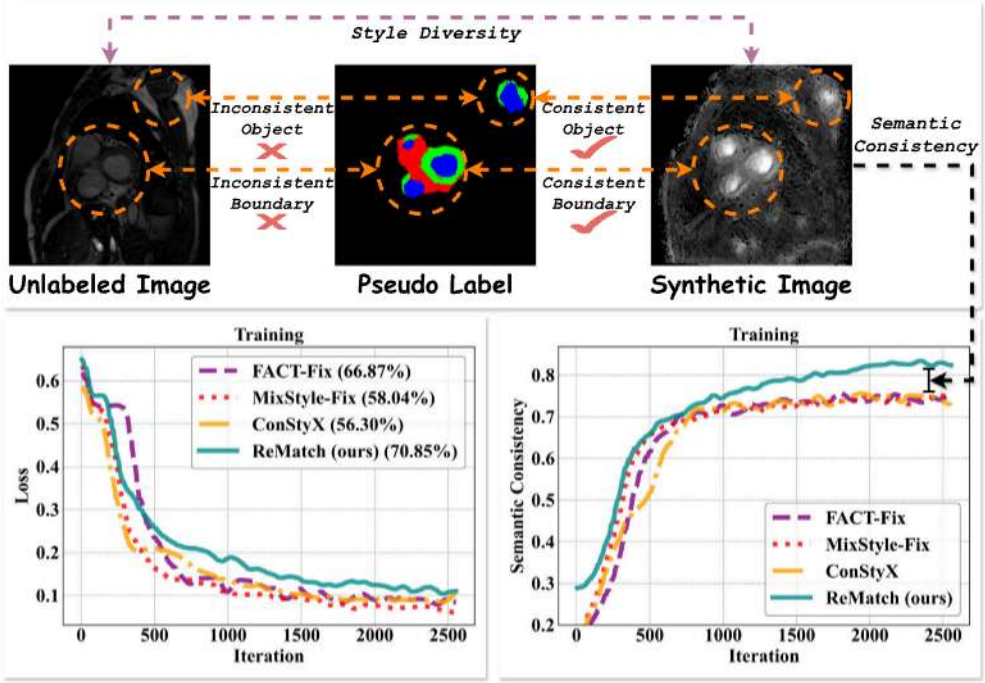


Fig. 1. Top: Illustration of *semantic consistency* and *style diversity* of the SynTS synthetic unlabeled-image and pseudo-label pairs. Refer to Fig. 6 for additional examples. Bottom: Training loss curves (along with cross-domain *test* performance measured by Dice similarity coefficients) and semantic consistency on unlabeled data between predictive labels and ground truth (measured by Dice similarity coefficients) for FACT-Fix [3], MixStyle-Fix [4], ConStyX [5], and our ReMatch. Because both the SynTS synthesized image and its pseudo label are derived from the same segmentation model for an unlabeled image, they inherently encode consistent semantic information. Hence, the pseudo labels can be regarded as semantically consistent supervisory labels for the synthesized images.

frequency perturbation [10,11], feature statistics editing [5,12–14], and random convolution [15,16], focus on varying image styles and cannot modify the semantic information of unlabeled images to match their pseudo labels [5,10–13,15,16]. In contrast, generative adversarial networks or diffusion models-based image synthesis methods require large-scale training data to produce images with fine-grained visual details [17–20]. Such large-scale data are impractical in SSDG medical image segmentation scenarios. In light of the above, we pinpoint an unresolved problem: *How to ensure semantic consistency and style diversity of unlabeled-image and pseudo-label pairs for training a well-generalized model?*

To address this problem in SSDG medical image segmentation scenarios, we propose **ReMatch**, a segmentation-synthesis co-training framework that performs consistency regularization on the segmentation model using both original and synthetic images. At its core is **SynTS** that **Syn**thesizes unlabeled images with both high *semantic consistency (with pseudo labels)* and *style diversity* by leveraging **T**exture and **S**hape features extracted during the segmentation process. In the co-training process, the segmentation model provides intermediate features for image synthesis, and SynTS generates synthetic-image and pseudo-label pairs for model training, ultimately yielding a well-generalized segmentation model. Specifically, SynTS involves two steps: (1) image synthesis, exploiting the model prior for texture and shape feature extraction and the data prior for content and style modeling to generate style-content basis images, and (2) image transformation, transforming the style-content basis image domain to the real image domain. Because the synthetic images and their corresponding pseudo labels are derived from shape features extracted by the same model from identical unlabeled images, they share semantic information and thus ensure high semantic consistency. Meanwhile, the continuous updating of model parameters during training leads to variations in the extracted texture features; using these dynamic features as style components endows the synthetic images with high style diversity. These two properties of the synthetic image-label pairs provide high-quality pseudo labels for effective consistency regularization on the segmentation model and expand the source-domain data distribution, jointly addressing the challenges of label scarcity and domain shift in SSDG.

Our contributions can be summarized as follows:

- We identify the importance of simultaneously ensuring two key principles for SSDG medical image segmentation: (1) semantic consistency between unlabeled images and pseudo labels to extract reliable supervisory signals from unlabeled data, and (2) style diversity of unlabeled images to facilitate the training of well-generalized models.
- We propose the SynTS algorithm, which exploits the model and data priors to synthesize unlabeled-image and pseudo-label pairs with both high semantic consistency and style diversity without relying on large-scale data. Building upon the SynTS algorithm, we develop a new segmentation-synthesis co-training framework for SSDG medical image segmentation.
- Extensive experiments on cardiac magnetic resonance images (MRI) and fundus photography datasets with single-source single-target and single-source multi-target cross-domain settings, respectively, demonstrate that ReMatch is a competitive alternative to the state-of-the-art approaches.

2. Related work

2.1. Semi-supervised learning

Given the in-domain assumption for the training and test sets, semi-supervised learning (SSL) aims to train a model using limited labeled data and numerous unlabeled images to alleviate label scarcity [21]. This line of research evolved from cardinal consistency regularization [22] and pseudo-labeling [23] to a more effective learning paradigm: strong–weak pseudo-supervision [9,24]. Recently, the concept of strong–weak pseudo-supervision has been widely applied to

semantic segmentation tasks in both natural [25–29] and medical [30–37] images, leading to a series of state-of-the-art approaches. For example, some effective ideas have emerged to exploit unlabeled data, including a latent feature rectification [32], a pseudo-label-guided masking and reconstruction scheme [34], and an adaptive Mixup perturbation [35], *etc*. However, domain shifts arising from variations in patient populations, imaging devices, and operator expertise limit the underlying assumption of SSL in clinical applications. This paper investigates a more challenging yet practical problem: semi-supervised domain generalization.

2.2. Domain generalization

Under the out-of-domain assumption between the training and test sets, domain generalization (DG) methods train a model on a fully annotated dataset derived from the source domain and aim to generalize it to an unseen domain [38]. Data augmentation is the mainstream strategy for DG, facilitating the model to learn domain-invariant representations by expanding the source domain data distribution (*a.k.a.*, domain randomization) [5,10–13,15,16,39–43]. Specifically, this class of techniques mainly includes: (1) frequency transformation-based methods, which combine Fourier amplitude spectra from different images for style augmentation [10,11,39,40,44]; (2) feature statistics editing-based approaches, which perturb image- or feature-level statistics (*e.g.*, mean and standard deviation) to achieve image style or content augmentation [5,12–14]; and (3) random convolution-based methods, which non-linearly perturb image intensity using convolutional neural networks with random parameters [15,16,42,43]. However, the fully annotated setting increases annotation costs and limits DG methods from leveraging unlabeled data. This work goes beyond the DG setting and investigates the more challenging problem of semi-supervised domain generalization.

2.3. Semi-supervised domain generalization

Under the partially annotated training set and cross-domain generalization settings, semi-supervised domain generalization (SSDG) aims to address both label scarcity and domain shifts, and can be regarded as a combination of SSL and DG tasks. Current SSDG approaches have primarily been built upon the strong–weak pseudo supervision paradigm to leverage unlabeled images, along with specific data augmentation techniques to expand the style diversity of source domain data [3,4,6–8,45,46]. Representative methods include co-training or self-training frameworks equipped with frequency augmentation [3,6], extended FixMatch frameworks with MixStyle [4], and an asymmetric co-training scheme with frequency augmentation [46]. Despite their favorable results, semantic inconsistencies between style-augmented unlabeled images and their pseudo labels undermine the generalizability of trained models, especially in the field of medical images, where generalization ability is a critical concern for clinical applications involving life-threatening risks. To address this problem, we design the SynTS algorithm, which ensures both semantic consistency and style diversity in the synthetic-image and pseudo-label pairs, enabling the training of a well-generalized model from a partially annotated dataset for cross-domain medical image segmentation.

3. Method

Before presenting the proposed method, we introduce the notations used in this work. The training set $\mathcal{D} = \{\mathcal{D}^L, \mathcal{D}^U\}$ includes a labeled set $\mathcal{D}^L = \{S_k | S_k = \{(x^l_{i(k)}, y^l_{i(k)})_{i=1}^{N_k^L}\}_{k=1}^{K}\}$ and an unlabeled set $\mathcal{D}^U = \{S_k | S_k = \{(x^u_{j(k)})_{j=1}^{N_k^U}\}_{k=1}^{K}\}$ with K domains, where $x^l_{i(k)}/x^u_{j(k)}$ represents the ith/jth labeled/unlabeled image in the k^{th} domain and $y^l_{i(k)}$ is the ground truth for $x^l_{i(k)}$. N_k^L and N_k^U ($N_k^U \gg N_k^L$) indicate the numbers of labeled and unlabeled samples in the kth domain. Note that we omit the domain index k for brevity, as the training data involves only one domain, *i.e.*, $K = 1$. Given the training data $\mathcal{D}$, semi-supervised domain generalizable segmentation aims to learn a model $f(\cdot;\theta)$ that performs well on a test set $\mathcal{T}$ from one or more unseen target domains.

3.1. Preliminaries

Strong–Weak Pseudo Supervision Given $(x_i^l, y_i^l) \in \mathcal{D}^L$, $x_j^u \in \mathcal{D}^U$, and its corresponding pseudo label $\bar{y}_j^u$ with the confidence map $\bar{p}_j^u$, the objective for strong–weak pseudo supervision can be formulated as:

$$J = \min_{\theta} \mathbb{E}_{\mathcal{D}}[\underbrace{\mathcal{L}(f(\Gamma(x_i^l),\theta), y_i^l)}_{\mathcal{L}_s} + \lambda \underbrace{\mathbb{I}_{\{\bar{p}_j^u \geq \tau\}}(\bar{p}_j^u)\mathcal{L}(f(\Gamma(x_j^u),\theta), \bar{y}_j^u)}_{\mathcal{L}_u}] \tag{1}$$

where $\mathcal{L}_s$ and $\mathcal{L}_u$ denote the supervised and unsupervised losses, respectively. λ is a weighting parameter to control the contribution of the unsupervised loss. $\mathcal{L}(\cdot,\cdot)$ refers to a criterion for the underlying task, *e.g.*, image recognition or semantic segmentation, $f(\cdot,\theta)$ denotes a model with learnable parameters θ, and τ is a confidence threshold above which a pixel-level pseudo label is selected for pseudo supervision. $\Gamma(\cdot)$ indicates a strong augmentation. The pseudo label $\bar{y}_j^u$ can be generated through $\bar{y}_j^u = \arg\max \mathrm{SoftMax}(f(\gamma(x_j^u),\theta))$, where the specific implementation depends on the chosen learning paradigm (*e.g.*, self-training [9], mean teacher [22], co-training [25], or more complicated one). Here, $\gamma(\cdot)$ is weak augmentation, *w.r.t.*, the strong one. In general, $\gamma(\cdot)$ only performs geometric transformations (*e.g.*, rotation/scaling) on input images, whereas $\Gamma(\cdot)$, in addition to $\gamma(\cdot)$, incorporates more sophisticated augmentations, including pixel intensity augmentations (*e.g.*, brightness/contrast variations) and Mixup perturbations (*e.g.*, CutMix [47]) [9].

3.2. Rematch framework

As illustrated in Fig. 2, ReMatch is a segmentation-synthesis co-training framework including two main components: (1) the SynTS algorithm SynTS$(\cdot,\phi)$, parameterized by ϕ, which synthesizes unlabeled-image and pseudo-label pairs using the texture and shape features provided by the segmentation model, (2) the segmentation model $f(\cdot,\theta)$ that is trained via a supervised learning path on labeled data and an unsupervised learning path on both the original and synthetic unlabeled data.

Formally, the training loss for the ReMatch framework involves three parts: (1) the supervised loss $\mathcal{L}_s$ on labeled data, (2) the unsupervised loss $\mathcal{L}_u$ on both original and synthetic unlabeled images, and (3) the reconstruction loss $\mathcal{L}_r$ for training the image transformation model within the SynTS algorithm, which is specified in Section 3.3.

$$\mathcal{L} = \mathcal{L}_s + \lambda\mathcal{L}_u + \eta\mathcal{L}_r, \tag{2}$$

where λ and η denote weighting parameters that control the contributions of the unsupervised loss and the reconstruction loss, respectively.

Supervised learning path. For labeled data, the supervised loss $\mathcal{L}_s$ is formulated as:

$$\mathcal{L}_s = \mathcal{L}_{seg}(f(\Gamma(x_i^l),\theta), y_i^l) \tag{3}$$

where $\mathcal{L}_{seg}$ refers to a segmentation criterion.

Unsupervised learning path. Based on the strong–weak pseudo supervision in Eq. (1), the proposed "*rematch*" unsupervised loss $\mathcal{L}_u$ consists of two terms that extract supervisory information from both the original unlabeled images and the SynTS synthetic unlabeled images:

$$\mathcal{L}_u = \underbrace{\mathbb{I}_{\{\bar{p}_j^u \geq \tau\}}(\bar{p}_j^u)\mathcal{L}(f(\Gamma(x_j^u),\theta), \bar{y}_j^u)}_{\mathcal{L}_{uo}\text{ for original unlabeled image}} + \underbrace{\mathbb{I}_{\{\bar{p}_j^u \geq \tau\}}(\bar{p}_j^u)\mathcal{L}(f(\tilde{x}_j^u,\theta), \bar{y}_j^u)}_{\mathcal{L}_{us}\text{ for synthetic unlabeled image}}, \tag{4}$$

where $\tilde{x}_j^u$ represents a SynTS synthetic unlabeled image. We retain the unsupervised loss term on the original unlabeled images as a

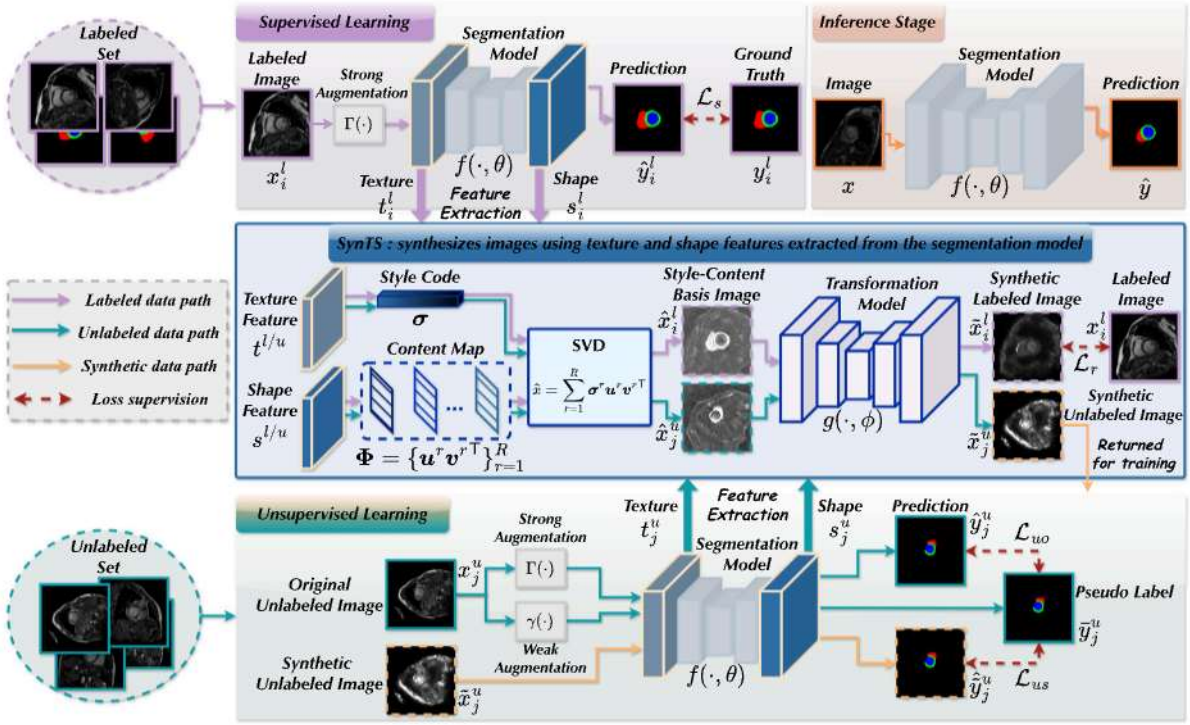


Fig. 2. Illustration of the ReMatch framework. ReMatch is a segmentation-synthesis co-training framework built upon the SynTS algorithm for semi-supervised domain generalizable medical image segmentation. Note that only the segmentation model is required in the inference stage.

compromise to regularize the model on the source domain and prevent overfitting to the synthetic domain (as illustrated in Fig. 6, the synthetic images differ significantly from the original ones). Meanwhile, although each synthetic image is well-aligned with its pseudo label, we further apply a thresholding strategy to ensure that the segmentation targets in the pseudo label do not deviate significantly from real anatomical structures.

3.3. SynTS algorithm

To ensure controllability of the style and content of synthetic images and to mitigate reliance on large-scale data, we devise the SynTS algorithm, which synthesizes augmented unlabeled images with both high *semantic consistency* with pseudo labels and *style diversity* by exploiting the texture and shape features extracted during the segmentation process. As illustrated in Fig. 2, the SynTS algorithm consists of two main steps: (1) image synthesis, which generates style-content basis images; and (2) image transformation, which transforms the style-content basis image domain to the real image domain.

3.3.1. Image synthesis

This step explores prior knowledge at both the *model* and *data* levels to achieve adaptive control over the style and content of synthetic images and to reduce reliance on large-scale data.

Model prior. In convolutional neural networks, shallow layers predominantly capture low-level **texture** features (*e.g.*, local edges, corners, and color), whereas deep layers encode high-level **shape** information, such as the structure of target objects [1,48,49]. Building on this concept, SynTS utilizes feature maps from a shallow layer to capture texture features and from a deep layer to represent shape features.

Data prior. Images can be represented as matrices. As demonstrated by principal component analysis, the low-rank matrices of an image correspond to its principal components [50–52]. Building on this insight, our previous work [53] reveals that applying singular value decomposition (SVD) to an image effectively disentangles its representations: the singular values encode its **style codes** and the corresponding rank-one matrices capture its **content maps**:

$$x = \sum_{r=1}^{R} \underbrace{\sigma^r}_{\text{style code}} \underbrace{\boldsymbol{u}^r \boldsymbol{v}^{r\top}}_{\text{content map}} \tag{5}$$

where R denotes the rank of x, the singular value σ^r is a style code, and the rank-one matrix $\boldsymbol{u}^r \boldsymbol{v}^{r\top}$ acts as a content map.

Content Map. The set of rank-one matrices $\boldsymbol{\Phi} = \{\boldsymbol{u}^1\boldsymbol{v}^{1\top}, \boldsymbol{u}^2\boldsymbol{v}^{2\top}, \ldots, \boldsymbol{u}^R\boldsymbol{v}^{R\top}\}$ constitutes the content maps of image x, reflecting its anatomical structure. SVD provides the optimal rank-one decomposition in the least-squares sense and guarantees that the set $\boldsymbol{\Phi}$ forms a complete basis for the column and row spaces of x. Each rank-one component in $\boldsymbol{\Phi}$ serves as a basis pattern of the image, and any possible image structure can be expressed as a linear combination of these patterns. Thus, $\boldsymbol{\Phi}$ completely defines the anatomical structures of the image.

Style Code. The singular value vector $\boldsymbol{\sigma}$ acts as the style code that modulates the magnitude of pixel intensity of image x. From the reconstruction formula in Eq. (5), each scalar σ^r acts as a global multiplier for the corresponding basis pattern ($\boldsymbol{u}^r\boldsymbol{v}^{r\top}$). Variations in σ^r uniformly influence the intensity of all pixels in the synthesized image x, proportionally scaling the values of the basis pattern.

Applying a singular value vector $\boldsymbol{\sigma}$ to the basis set $\boldsymbol{\Phi}$ yields an image in which overall contrast and visual style are governed by $\boldsymbol{\sigma}$, while anatomical content is determined by $\boldsymbol{\Phi}$. Fig. 3 shows a decomposition example on the balanced Steady-State Free Precession (bSSFP) and Late Gadolinium Enhancement (LGE) MRI sequences from the MS-CMR dataset [54]. It can be observed that exchanging the style codes of the two images enables style transfer between them, while summing all content maps of an image reconstructs its complete anatomical structure.

By leveraging the *model* and *data* priors, the synthetic style-content basis images have two major advantages:

- **Semantic Consistency:** Since pseudo labels rely on semantic information aggregated in a deep layer of the segmentation model,

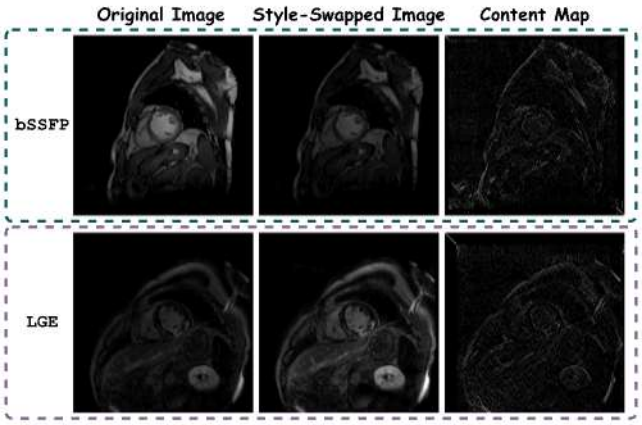


Fig. 3. Illustration of the SVD-based style-content decomposition using bSSFP and LGE MRI sequences from the MS-CMR dataset. Style-swapped images are generated by exchanging the style codes of the two original images (top: bSSFP image with LGE style; bottom: LGE image with bSSFP style). Content maps are obtained by summing the rank-one matrices of each image.

their content characteristics align more closely with the shape feature maps than with the input images. Consequently, compared with original unlabeled images, the synthetic images generated from content maps extracted from the shape features exhibit higher semantic consistency with the pseudo labels.

- **Style Diversity:** Texture features are extracted from a shallow layer whose parameters are gradually updated during training. These updates introduce variations in the style codes derived from the texture features, resulting in high style diversity of the synthesized images.

As illustrated in Fig. 2, we first extract the style code $\boldsymbol{\sigma}$ and a set of content maps $\boldsymbol{\Phi}$ from the texture and shape feature maps, respectively, for both labeled and unlabeled images using Eq (5).

$$\begin{aligned} &t_j^u \rightarrow \boldsymbol{\sigma}_j^u \text{ and } s_j^u \rightarrow \{\boldsymbol{\Phi}_j^{u(r)}\}_{r=1}^{R_j}, \text{ for an unlabeled image} \\ &t_i^l \rightarrow \boldsymbol{\sigma}_i^l \text{ and } s_i^l \rightarrow \{\boldsymbol{\Phi}_i^{l(r)}\}_{r=1}^{R_i}, \text{ for a labeled image} \end{aligned} \tag{6}$$

Then, the synthetic unlabeled and labeled images, $\hat{x}^u$ and $\hat{x}^l$, are reconstructed via SVD:

$$\hat{x}_j^u = \sum_{r=1}^{R_j} \sigma_j^{u(r)} \boldsymbol{\Phi}_j^{u(r)}, \quad \hat{x}_i^l = \sum_{r=1}^{R_i} \sigma_i^{l(r)} \boldsymbol{\Phi}_i^{l(r)} \tag{7}$$

3.3.2. Image transformation

As shown in Fig. 2, the synthetic style-content basis images deviate from real data due to the effects of convolution layers. To map the style-content basis image domain to the real image domain, we employ a transformation model $g(\cdot, \phi)$ with trainable parameters ϕ on the synthetic images:

$$\begin{aligned} &\tilde{x}_j^u = g(\hat{x}_j^u, \phi), \text{ (without gradient)} \\ &\tilde{x}_i^l = g(\hat{x}_i^l, \phi), \text{ (with gradient)} \end{aligned} \tag{8}$$

where $\tilde{x}_j^u$ and $\tilde{x}_i^l$ denote the final synthetic unlabeled and labeled images, respectively.

The training of the transformation model relies solely on labeled data, due to the following reasons: (1) If unlabeled data are used for training, the trained transformation model tends to reconstruct the original unlabeled images from the synthesized ones, making the image synthesis process unavailing. (2) For labeled data, the supervision from ground truth labels ensures that the shape features are consistent with the original image structure. This allows the transformation model to concentrate on enhancing image quality without altering the underlying image content. Consequently, the reconstruction loss $\mathcal{L}_r$ is defined as:

$$\mathcal{L}_r = \mathcal{L}_{rec}(g(\hat{x}_i^l, \phi), x_i^l) \tag{9}$$

where $\mathcal{L}_{rec}$ refers to an image reconstruction criterion.

4. Experiments and results

4.1. Tasks & datasets

Cross-domain cardiac segmentation on the MS-CMR dataset (single source and single target domain). The MS-CMR dataset[1] [54] contains 45 subjects, each with bSSFP and LGE MRI sequences, along with ground truth annotations for the right ventricle (RV), left ventricle (LV), and myocardium (MYO). All images are normalized to the range $[0, 1]$ and resampled to a uniform resolution of 1.0×1.0 mm. This experimental setting evaluates a segmentation model's ability to generalize across MRI sequences for cardiac structure segmentation, specifically from LGE to bSSFP and vice versa. For each direction, one sequence (*e.g.*, bSSFP) is treated as the source domain and is divided into training and validation sets with an $8:2$ ratio, while the other sequence (*e.g.*, LGE) serves as the target domain and is used solely for testing.

Cross-domain Optic Cup (OC) and Optic Disc (OD) segmentation on the fundus image benchmark (single source and multiple target domains). The benchmark[2] [13] is established based on the ORIGA [55], Drishti-GS [56], REFUGE [57], and RIGA [58] (BinRushed and Magrabia) datasets. This setting evaluates the model's generalization capability for the joint OC and OD segmentation across multiple target domains. The ORIGA dataset is used as the source domain with training and validation sets containing 500 and 150 images, respectively, following the data split protocol proposed by TriD [13]; the remaining datasets, including BinRushed (195 images), Drishti-GS (101 images), Magrabia (95 images), and REFUGE (400 images), serve as target domains for cross-domain evaluation.

4.2. Implementation details

Environmental setup. We conducted our experiments using an NVIDIA A40 GPU with 48 GB memory, Python 3.8, PyTorch 1.11.0, and CUDA 11.3. We employed the AdamW optimizer [59] with a fixed learning rate of $1e{-}4$ and a batch size of 16. All methods were trained for 100 epochs in all experiments.

Framework. We adopted the U-Net [1] architecture for both the segmentation baseline $f(\cdot, \theta)$ and the transformation model $g(\cdot, \phi)$. We extracted the feature maps from the first convolutional block of U-Net as texture features, and those from the final convolutional block (before the segmentation head) as shape features. Based on the hyperparameter study, we set the confidence threshold to $\tau = 0.95$, the unsupervised loss weight to $\lambda = 1$, and the reconstruction loss weight to $\eta = 10$. The weak augmentation function $\gamma(\cdot)$ includes random flipping and rotation; building upon $\gamma(\cdot)$, strong augmentation $\Gamma(\cdot)$ further incorporates ColorJitter, Gaussian blur, CutMix [47], and StyCona [53]. For cardiac MRI, the SynTS algorithm directly synthesizes the image itself; for fundus images with three channels, SynTS synthesizes the Y channel and combines it with the original image's Cb and Cr channels to generate the synthetic image. For the loss functions, we utilized the Dice loss [2] as the segmentation criterion and the Mean Squared Error (MSE) loss as the reconstruction criterion. All images were resized to 256×256 for both training and testing. During evaluation, the predicted labels were rescaled to their original resolutions.

[1] https://zmiclab.github.io/zxh/0/mscmrseg19/index.html.
[2] https://zenodo.org/records/8009107.

Table 1
Comparison with state-of-the-art methods on the cardiac segmentation task with **10%** and **20%** labeled data. Note that all the compared DG methods are equipped with the strong–weak pseudo supervision paradigm for using unlabeled data. The best and second-best results are highlighted in **bold** and underline, respectively.

Method	bSSFP → LGE				LGE → bSSFP				Average			
	10% $\|D\|$		20% $\|D\|$		10% $\|D\|$		20% $\|D\|$		10% $\|D\|$		20% $\|D\|$	
	DSC (%) ↑	ASD ↓	DSC (%) ↑	ASD ↓	DSC (%) ↑	ASD ↓	DSC (%) ↑	ASD ↓	DSC (%) ↑	ASD ↓	DSC (%) ↑	ASD ↓
AmpSwap [10]	68.47	4.31	68.51	4.36	77.65	**3.01**	76.81	3.83	73.06	3.66	72.66	4.10
EPLSDG [6]	50.23	5.15	57.86	6.48	59.41	4.45	77.35	2.63	54.84	4.80	67.61	4.56
FACT-Fix [3]	66.87	4.73	69.04	4.92	77.84	3.10	78.59	2.25	72.36	3.92	73.82	3.59
TriD [13]	60.94	6.60	66.96	15.76	76.73	3.89	77.53	3.53	68.84	5.25	72.25	9.65
MixStyle-Fix [4]	58.04	6.42	65.13	8.86	76.00	4.81	76.32	5.33	67.02	5.62	70.73	7.10
ConStyX [5]	56.30	6.95	59.99	4.78	77.07	3.88	78.61	2.59	66.69	5.42	69.30	3.69
RandConv [15]	64.87	9.31	69.54	6.94	76.84	3.96	76.43	3.37	70.86	6.64	72.99	5.16
CIDA [42]	**71.99**	9.75	**73.25**	6.78	71.05	7.58	73.62	4.43	71.52	8.67	73.44	5.61
PRandConv [16]	66.53	7.29	66.12	11.52	75.00	3.97	74.02	3.86	70.77	5.63	70.07	7.69
ReMatch (ours)	70.85	**3.41**	71.73	**3.08**	**78.48**	3.32	**79.86**	**2.02**	**74.67**	**3.37**	**75.80**	**2.55**

Table 2
Comparison with state-of-the-art methods on the fundus segmentation tasks under **10%** and **20%** labeled data (Source: ORIGA; Target: REFUGE, Drishti-GS, BinRushed, and Magrabia). Note that all the compared DG methods are equipped with the strong–weak pseudo supervision paradigm for using unlabeled data. Results are measured by DSC. The best and second-best results are highlighted in **bold** and underline, respectively.

Method	BinRushed		Drishti-GS		Magrabia		REFUGE		Average	
	10% $\|D\|$	20% $\|D\|$	10% $\|D\|$	20% $\|D\|$	10% $\|D\|$	20% $\|D\|$	10% $\|D\|$	20% $\|D\|$	10% $\|D\|$	20% $\|D\|$
AmpSwap [10]	59.01	59.72	78.51	76.54	60.15	58.34	80.42	79.94	69.55	68.64
EPLSDG [6]	**67.76**	69.60	72.72	72.11	**67.23**	**71.80**	78.25	77.31	71.49	72.71
FACT-Fix [3]	64.04	67.73	77.60	78.06	64.74	64.26	79.76	81.92	71.54	72.99
TriD [13]	62.82	69.02	77.73	77.28	63.76	65.23	**81.66**	81.98	71.49	73.38
MixStyle-Fix [4]	64.62	70.21	79.68	74.43	66.64	65.58	80.35	80.79	72.82	73.25
ConStyX [5]	60.86	68.84	80.37	**78.72**	60.87	64.11	80.64	82.77	70.69	73.61
RandConv [15]	55.21	63.39	79.55	78.49	63.01	67.22	80.32	80.48	69.52	72.40
CIDA [42]	48.95	48.37	75.12	76.67	51.17	52.46	69.09	66.74	61.08	61.06
PRandConv [16]	43.32	37.62	71.66	78.56	54.59	47.83	77.40	74.98	61.74	59.75
ReMatch (ours)	65.37	**72.46**	**80.41**	78.25	65.56	64.43	81.60	**82.95**	**73.22**	**74.52**

Evaluation metrics. Dice Similarity Coefficient (DSC) and Average Surface Distance (ASD) were used to evaluate segmentation performance in terms of spatial overlap and surface topology, respectively.

4.3. Comparison with state of the art

We compared ReMatch with state-of-the-art DG and SSDG methods, which can be divided into three groups based on their data augmentation strategies: (1) frequency transformation-based: AmpSwap [10], EPLSDG [6], and FACT-Fix [3]; (2) feature statistics editing-based: TriD [13], MixStyle-Fix [4], and ConStyX [5]; (3) random convolution-based: RanConv [15], CIDA [42], and PRanConv [16]; Note that, for a fair comparison, all the compared DG methods (AmpSwap [10], TriD [13], ConStyX [5], RanConv [15], CIDA [42], and PRanConv [16]) are equipped with the strong–weak pseudo-supervision paradigm [9] for utilizing unlabeled data.

4.3.1. Results on cardiac segmentation

In this experiment, we evaluated the effectiveness of ReMatch in a single-source single-target cross-domain medical image segmentation setting. As shown in Table 1, ReMatch shows superior performance over the compared methods in both the bSSFP → LGE and LGE → bSSFP scenarios, under the settings with 10% and 20% labeled data. The feature statistics editing-based methods present relatively inferior performance compared with the other classes of approaches, as the style transfer mechanism is less effective for grayscale images. Notably, CIDA [42], an improved random convolution-based method, achieves the highest Dice scores in the bSSFP → LGE scenario but suffers a dramatic performance drop in the LGE → bSSFP scenario. This discrepancy may result from the global random convolution and local intervention augmentations in CIDA. While these operations generate augmented images from the bSSFP modality that are beneficial for LGE domain segmentation, they suppress segmentation-relevant information when generalizing from the LGE to the bSSFP domain. In contrast, ReMatch obtains more robust performance across these scenarios, particularly on the boundary-sensitive metric (*i.e.*, ASD) owing to the stable training enabled by the consistent synthetic-image and pseudo-label pairs.

4.3.2. Results on Optic Disc and Optic Cup segmentation

This experiment further evaluated the performance of ReMatch in a more challenging single-source multiple-target cross-domain medical image segmentation setting. Table 2 reports the average results of the joint OC and OD segmentation when ORIGA is used as the source domain with 10% and 20% labeled data, respectively, and the other datasets serve as the multiple target domains. Generally, our ReMatch achieves the best average DSC over the compared methods, especially those (*i.e.*, TriD [13], MixStyle-Fix [4], and ConStyX [5]) with style transfer that are more suitable for RGB images. Fourier transformation-based methods demonstrate moderate performance, consistent with the trends observed in the cardiac segmentation task. In contrast, random convolution-based methods deliver suboptimal performance, which may be attributed to improper distortions introduced by random nonlinear augmentations on pixel intensities. The results suggest that ReMatch can serve as a robust solution for a wide range of cross-domain medical image segmentation tasks, highlighting its versatility across different imaging modalities.

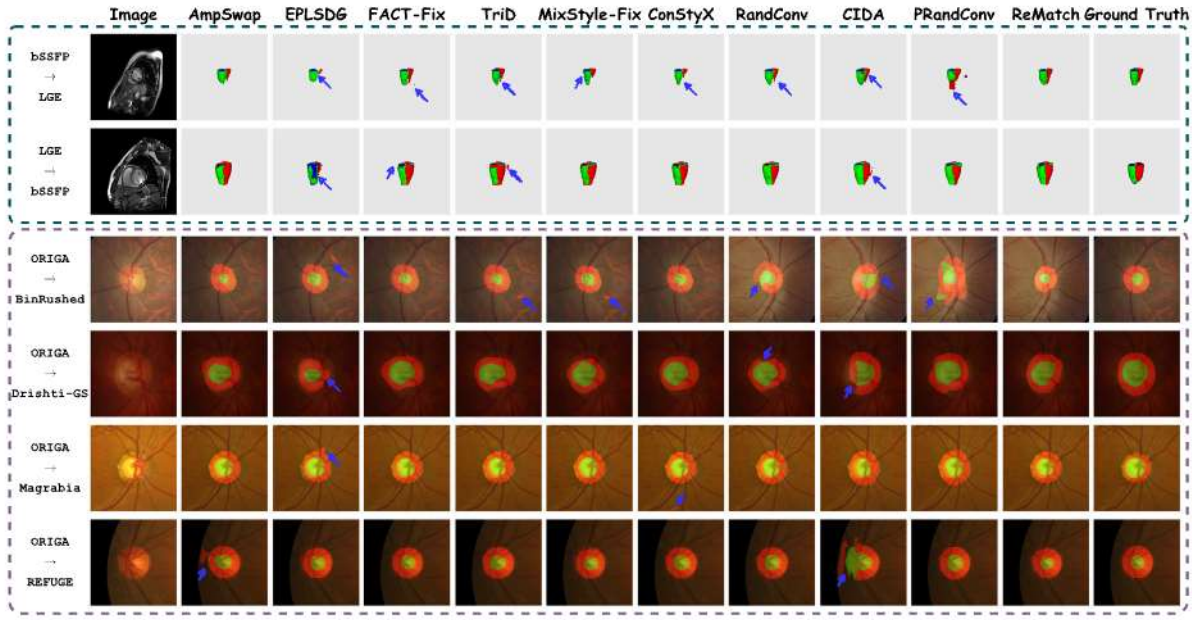


Fig. 4. Qualitative examples from the cardiac MRI and fundus image segmentation tasks. Blue arrows indicate representative segmentation errors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Ablation study for the ReMatch framework on the MS-CMR dataset with 10% and 20% labeled data. The best results are highlighted in **bold**.

ReMatch framework			bSSFP → LGE				LGE → bSSFP			
Strong–weak	SynTS algorithm		10% $\|D\|$		20% $\|D\|$		10% $\|D\|$		20% $\|D\|$	
pseudo supervision	Image synthesis	Image transformation	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓
			29.91	41.13	50.81	15.63	70.53	10.27	77.25	3.58
√			53.31	12.88	59.88	5.51	73.58	5.84	76.45	3.60
√	√		62.68	6.13	66.65	6.55	76.97	3.45	78.15	3.42
√	√	√	**70.85**	**3.41**	**71.73**	**3.08**	**78.48**	**3.32**	**79.86**	**2.02**

4.3.3. Qualitative results

In Fig. 4, we visualize some segmentation examples on the cardiac segmentation and the joint OC/OD segmentation tasks, respectively. The blue arrows highlight segmentation outliers, indicating that the corresponding models are overwhelmed by strong augmentations applied to images with low-quality pseudo labels. Overall, the qualitative examples are consistent with the quantitative results in Tables 1 and 2, further supporting our analysis. **Remark.** According to the quantitative and qualitative results, in general, the feature statistics editing-based methods perform relatively poorly on cardiac MRI images, mainly due to the lack of "color" variations in the gray-scale medical images, and these methods, in contrast, achieve better performance on fundus images. On the other hand, the random convolution-based methods (*e.g.*, CIDA) benefit from their nonlinear transformations of pixel intensities, yielding the best results in a specific cross-domain generalization scenario, *i.e.*, the bSSFP → LGE scenario. However, their performance on fundus images is unsatisfactory due to limited generalizability across multiple target domains. In addition, the Fourier transformation-based methods demonstrate balanced performance across both tasks, but the semantic inconsistencies between augmented images and pseudo labels constrain their effectiveness. In comparison, our ReMatch leverages the collaboration between image synthesis and segmentation to generate well-augmented images and corresponding high-quality pseudo labels, achieving competitive results on both tasks.

4.4. Ablation study

We conducted a series of ablation studies on the MS-CMR datasets with 10% and 20% labeled data to investigate: (1) the effectiveness of each component in the ReMatch framework [Table 3], (2) the impact of different hyperparameter choices [Table 4], (3) the influence of various image synthesis strategies [Fig. 5], and (4) visualization of the synthetic images during training [Fig. 6].

4.4.1. Effectiveness of each component

In this experiment, we investigated the effectiveness of both the strong–weak pseudo-supervision paradigm and the SynTS algorithm within the proposed ReMatch framework. As shown in Table 3, performance gradually improves as each component is progressively incorporated into ReMatch. The improvements are more significant in the bSSFP → LGE scenario, which may be attributed to the limited information in the bSSFP sequence for generalizing to the LGE sequence. Remarkably, introducing unlabeled images into model training leads to a decline in segmentation performance in the LGE → bSSFP setting with 20% labeled data. This phenomenon can be attributed to the model overfitting to the unlabeled source domain data, especially those associated with low-quality pseudo labels. In contrast, benefiting from the synthetic-image and pseudo-label pairs with high semantic consistency and style diversity, our final method, ReMatch, achieves the best performance in terms of both DSC and ASD metrics across all scenarios and labeled data rates.

4.4.2. Analysis of hyperparameters

Reconstruction loss weight η. In Table 4, we examined the impact of the weighting parameter η of $\mathcal{L}_r$ on ReMatch, using the MS-CMR dataset with 10% and 20% labeled data. ReMatch with $\eta = 10$ consistently outperforms other configurations, except for an outlier in ASD when $\eta = 1$. This result arises because scaling $\mathcal{L}_r$ by 10× aligns its magnitude

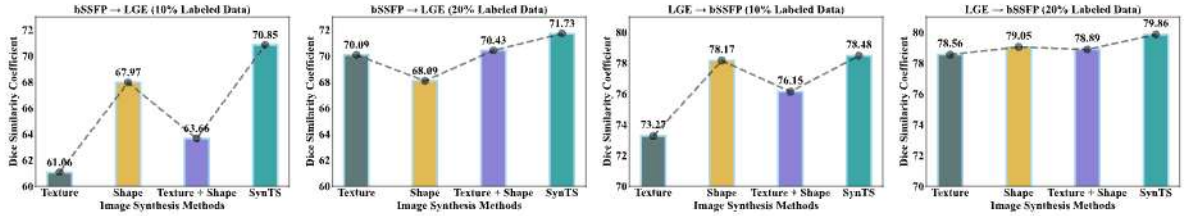


Fig. 5. Analysis of image synthesis strategies for the ReMatch framework on the MS-CMR dataset with 10% and 20% labeled data, respectively.

Table 4
Investigation of the impact of hyperparameters on the ReMatch framework using the MS-CMR dataset with 10% and 20% labeled data. The best results are highlighted in **bold**.

Hyperparameters	bSSFP → LGE				LGE → bSSFP			
	10% $\|D\|$		20% $\|D\|$		10% $\|D\|$		20% $\|D\|$	
	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓
$\eta = 1$	64.99	**2.21**	66.27	**3.03**	76.77	3.40	78.40	3.24
$\eta = 5$	68.88	5.31	70.19	3.73	78.02	3.46	76.55	3.58
$\eta = 10$	**70.85**	3.41	**71.73**	3.08	**78.48**	**3.32**	**79.86**	**2.02**
$\eta = 20$	69.58	4.30	65.47	5.03	76.11	3.37	76.51	3.88
$\lambda = 0.5$	68.56	**3.40**	70.48	3.79	77.82	3.78	78.37	4.32
$\lambda = 1$	**70.85**	3.41	**71.73**	3.08	**78.48**	**3.32**	**79.86**	**2.02**
$\lambda = 2$	70.18	4.87	71.66	**2.57**	76.72	3.48	78.89	2.61
$\tau = 0.9$	68.33	4.07	69.56	3.23	77.71	**3.04**	79.09	3.43
$\tau = 0.95$	**70.85**	**3.41**	**71.73**	**3.08**	**78.48**	3.32	**79.86**	**2.02**
$\tau = 0.99$	66.06	3.74	67.08	6.73	73.89	3.98	78.34	3.25
$\mathcal{L}_r = L_p$	66.79	4.63	70.26	4.21	**78.57**	**3.26**	**79.98**	2.60
$\mathcal{L}_r = L_2$	**70.85**	**3.41**	**71.73**	**3.08**	78.48	3.32	79.86	**2.02**

with those of the supervised and unsupervised segmentation losses, facilitating balanced optimization among these loss components during training.

Unsupervised loss weight λ. We investigated the sensitivity of ReMatch to the unsupervised loss weight by evaluating $\lambda \in \{0.5, 1, 2\}$. As shown in Table 4, $\lambda = 1$ delivers the best performance in all scenarios. A lower weight ($\lambda = 0.5$) weakens the contribution of unlabeled data, limiting the model to learn domain-invariant representations. In contrast, a higher weight ($\lambda = 2$) overemphasizes the unsupervised objective, which may amplify errors in pseudo labels and diminish the influence of reliable ground truth supervision from the limited labeled data. These results suggest that $\lambda = 1$ provides an appropriate balance between supervised and unsupervised learning.

Confidence threshold τ. To analyze the effect of pseudo label thresholding, we varied the confidence threshold $\tau \in \{0.9, 0.95, 0.99\}$. Table 4 indicates that $\tau = 0.95$ yields the highest DSC. A lower threshold ($\tau = 0.9$) allows low-confidence, noisy pseudo labels to participate in the co-training process, which degrades the effectiveness of consistency regularization from unlabeled data. On the other hand, a higher threshold ($\tau = 0.99$) discards a substantial portion of unlabeled and synthetic data, reducing the diversity of unlabeled data and hindering the model's convergence.

Reconstruction loss $\mathcal{L}_r$. Given that the MSE loss L_2 may cause the distortion of details and lead to blurry images, we explored the perceptual loss L_p [60] for enhancing the quality of synthetic images. Following [60,61], we employed the widely used pretrained VGG16 [62] to extract high-level features for the perceptual loss. As reported in Table 4, ReMatch with L_p obtains a noticeable performance drop in terms of DSC in the bSSFP → LGE scenario, while yielding only marginal gains in the LGE → bSSFP scenario. Although perceptual loss may generate higher-quality images, the high-level features it extracts tend to overemphasize semantic information, leading to exaggerated foreground regions and consequently causing inconsistencies between synthetic images and pseudo labels. These effects hinder cross-domain generalization. Moreover, the additional forward passes required for feature extraction in the perceptual loss substantially increase the computational overhead during training. In light of the above, we adopt MSE loss as the image reconstruction criterion in our method.

4.4.3. Analysis of image synthesis

To further analyze the role of image synthesis in ReMatch, we performed a series of experiments using the MSCMR dataset with 10% and 20% labeled data. As depicted in Fig. 5, we evaluated four image synthesis strategies: (1) using only texture features, (2) using only shape features, (3) combining texture and shape features via element-wise addition, and (4) the proposed SynTS algorithm. First, the texture-based synthesis method yielded the most unsatisfactory results overall. This can be attributed to two main factors: (i) texture features obtained from a shallow convolution layer often contain noise and artifacts caused by convolution operations, resulting in synthesized images that significantly deviate from real data; and (ii) semantic inconsistencies between the shape information embedded in the texture feature maps and the corresponding pseudo labels compromise the effectiveness of consistency regularization. In contrast, images synthesized from shape features tend to introduce spurious shortcut correlations between anatomical structures and segmentation targets that do not exist in the target domain, thereby impairing model generalization. Moreover, the fusion strategy fails to control the style and content of synthetic images, resulting in inconsistent synthetic pairs that hinder model training. SynTS addresses these limitations by separately modeling style and content: it extracts style codes from texture features and content maps from shape features to synthesize data with both high semantic consistency and style diversity. As a result, the model trained on the SynTS synthetic data exhibits superior generalization performance.

4.4.4. Visualization of SynTS synthetic images

To validate that SynTS can ensure both semantic consistency and style diversity of synthetic unlabeled-image and pseudo-label pairs, we visualized two examples from the cardiac and fundus segmentation

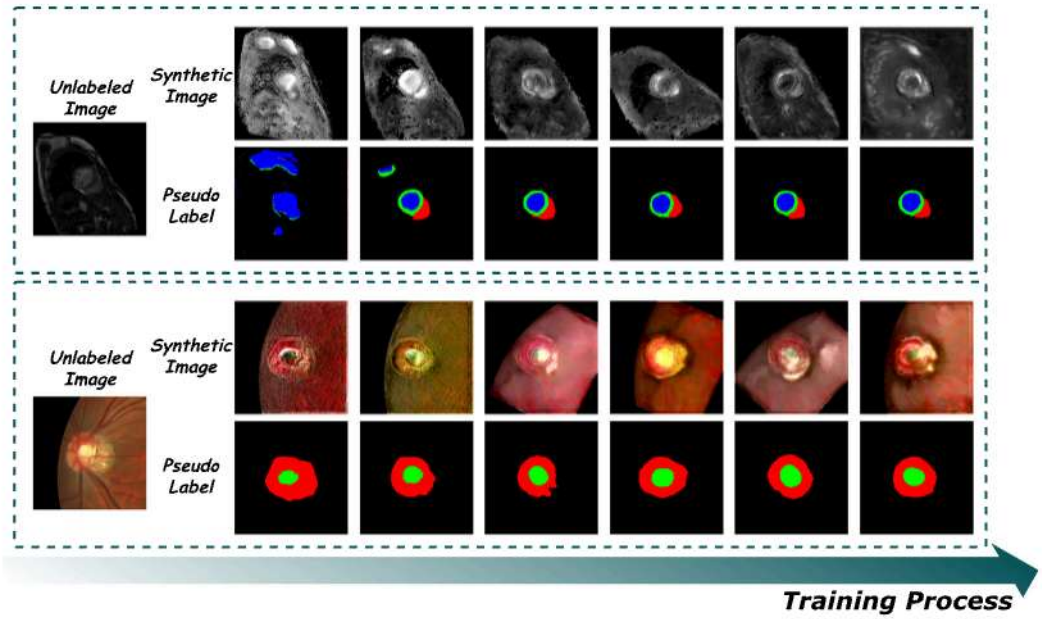


Fig. 6. Visualization of SynTS synthetic unlabeled-image and pseudo-label pairs during training from the cardiac and fundus segmentation tasks, respectively.

tasks during training. As shown in Fig. 6, the target anatomical structures in SynTS images align well with the pseudo labels throughout the training process. Meanwhile, the quality of pseudo labels gradually improves as training progresses (*i.e.*, their corresponding anatomical structures increasingly resemble the real ones). Moreover, the synthetic images exhibit diverse styles, derived from texture features extracted by the convolutional layer with gradually updated parameters. These properties of the SynTS synthetic data contribute to training a well-generalized segmentation model.

4.5. Framework flexibility analysis

In this section, we investigated the flexibility of the ReMatch framework by exploring its integration with three components: data augmentation, style-content modeling, and segmentation backbone.

4.5.1. Data augmentation

We further integrated ReMatch with two augmentation strategies: (1) Fourier amplitude Mixup (AmpMix) [11], and (2) random convolution-based augmentation (CIDA) [42]. As shown in Table 5, ReMatch achieves more comparable and stable results across the bSSFP → LGE and LGE → bSSFP scenarios. In contrast, ReMatch with AmpMix obtains slightly lower performance compared with ReMatch. Although ReMatch with CIDA achieves better performance in the bSSFP → LGE scenario, it experiences a performance drop in the LGE → bSSFP scenario. Nevertheless, all these approaches show performance improvements over their counterparts integrated with FixMatch, as reported in Table 1. These results demonstrate the potential of ReMatch as a generic framework that can be integrated with various data augmentation methods.

4.5.2. Style-content modeling

We further explored the flexibility of ReMatch by replacing its SVD-based style-content modeling with Adaptive Instance Normalization (AdaIN) [63], which considers feature statistics (mean and variance) as style components. As shown in Table 5, ReMatch (AdaIN) yields relatively inferior DSC and ASD scores, compared with the original ReMatch. This performance gap suggests that the framework is compatible with alternative style-content modeling strategies. Moreover, SVD guarantees a deterministic and orthogonal decomposition of style and content representations, enabling more robust disentanglement for image synthesis.

4.5.3. Segmentation backbone

To assess the architectural adaptability of ReMatch, we substituted the U-Net segmentation backbone with a Transformer-based architecture, *i.e.*, Swin-UNet [64]. As reported in Table 5, compared with the U-Net-based counterpart, ReMatch (Swin-UNet) presents superior performance in the bSSFP → LGE scenario, but shows a performance drop in the LGE → bSSFP scenario. The improvement can be attributed to Transformer's ability to capture long-range dependencies and global contextual information, which is effective for preserving anatomical structural integrity and enhancing boundary delineation. However, the powerful representational capacity may also render the model more susceptible to overfitting on source-domain data, resulting in cross-domain performance degradation in some scenarios. In summary, the experiment validates the compatibility of ReMatch with different segmentation architectures.

4.6. Computational complexity analysis

To evaluate the efficiency of ReMatch, we analyzed model parameters and floating-point operations (FLOPs) in Table 6. Note that all compared methods are built upon the U-Net segmentation backbone with approximately 1.8 M parameters and 3.0 GFLOPs per feedforward pass. Specifically, the training pipeline of ReMatch involves six feedforward passes (~18.0 GFLOPs) and four SVD operations (~0.8 GFLOPs, accounting for only 4.25% of the total computational cost). In comparison, FACT-Fix requires five feedforward passes (~15.0 GFLOPs) combined with Fourier augmentation and feature perturbations, respectively. MixStyle-Fix involves three feedforward passes (~9.0 GFLOPs) along with feature perturbation. ConStyX includes five forward passes (~15.0 GFLOPs) with feature perturbation. The additional forward passes and SVD computations required for image synthesis increase the computational cost of ReMatch during training. However, this additional training overhead leads to clear improvements in segmentation performance. For instance, it achieves average Dice score gains of 2.31%, 7.65%, and 7.98% over FACT-Fix [3], MixStyle-Fix [4], and ConStyX [5] respectively. More importantly, only the trained segmentation model is required during the inference phase, allowing ReMatch to operate with the same computational cost as a standard segmentation backbone.

Table 5
Analysis of the flexibility of the ReMatch framework with different strong augmentation strategies, style-content modeling methods, and segmentation backbones using the MS-CMR dataset with 10% and 20% labeled data. The best results are highlighted in **bold**.

Method	bSSFP → LGE				LGE → bSSFP			
	10% $\|D\|$		20% $\|D\|$		10% $\|D\|$		20% $\|D\|$	
	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓
ReMatch (AmpMix [11])	69.44	4.59	70.41	4.43	**78.93**	3.71	79.77	2.40
ReMatch (CIDA [42])	74.10	4.89	75.21	4.35	76.59	5.50	78.10	3.93
ReMatch (AdaIN [63])	67.66	3.72	70.48	5.38	77.00	3.59	77.98	4.21
ReMatch (Swin-UNet [64])	**75.52**	**2.40**	**77.57**	**1.86**	76.45	**1.59**	76.05	**1.74**
ReMatch	70.85	3.41	71.73	3.08	78.48	3.32	**79.86**	2.02

Table 6
Computational complexity analysis in terms of model parameters, FLOPs with 256 × 256 input, as well as average DSC on the MS-CMR dataset with 10% and 20% labeled data.

Method	Params (M)	FLOPs (G)	DSC (%): 10% $\|D\|$	DSC (%): 20% $\|D\|$%
FACT-Fix	1.8	15.0	72.36	73.82
MixStyle-Fix	1.8	9.0	67.02	70.73
ConStyX	1.8	15.0	66.69	69.30
ReMatch	3.6	18.8	74.67	75.80

5. Limitations and future work

Although ReMatch demonstrates competitive performance in semi-supervised domain generalization for medical image segmentation, it has several limitations that warrant further investigation.

Additional training computational overhead. While the inference phase requires the same computational cost as a standard segmentation backbone, the framework introduces additional computational and architectural overhead during training. Specifically, ReMatch involves performing SVD on feature maps and optimizing the auxiliary image transformation model $g(\cdot, \phi)$ for image synthesis in the training phase, as reported in Table 6. Future work could explore more efficient image synthesis methods [65,66] to reduce training cost and improve scalability.

Out-of-distribution generalization. Although ReMatch shows compelling performance on both single-source single-target [Table 1] and single-source multi-target [Table 2] cross-domain settings, its results remain unsatisfactory in highly varied out-of-distribution multi-target scenarios (*e.g.*, generalizing to BinRushed or Magrabia). To further alleviate these extreme domain shifts, future research could incorporate feature perturbation techniques [5,14] into the image synthesis pipeline to enhance the style diversity of synthetic images. Moreover, the applicability of ReMatch to a broader range of anatomical structures and imaging modalities (*e.g.*, cross-modality abdominal multi-organ segmentation [67]) remains to be investigated to establish its generalizability in more realistic clinical scenarios.

Style-content-controllable data generation. The current style-content modeling relies on SVD, supported by empirical validation from both our prior work [53] and the present study. Moving forward, we will explore more accurate and efficient style-content modeling to construct a style-content-controllable image-label generation framework to alleviate label scarcity in medical image analysis with limited annotations.

Synthetic image quality vs. cross-domain segmentation performance. Higher visual fidelity does not necessarily translate into better cross-domain segmentation performance. Instead, the effectiveness of synthesized image-label pairs for cross-domain segmentation may depend on a balance among semantic consistency, style diversity, and visual quality. Understanding the interplay among these factors will be an important direction for future research.

6. Conclusion

We propose the SynTS algorithm, which synthesizes unlabeled-image and pseudo-label pairs with both high semantic consistency and style diversity by leveraging texture and shape features extracted from the segmentation model. Building on SynTS, we develop an image segmentation-synthesis co-training framework that achieves competitive performance in both single-source single-target and single-source multiple-target semi-supervised domain generalizable medical image segmentation tasks. Experimental results also demonstrate that the proposed method functions as a flexible and generic framework that can be integrated with various data augmentation techniques, style-content modeling strategies, and segmentation backbones. Enhancing these components within the framework provides a potential direction for future research.

CRediT authorship contribution statement

Zhiqiang Shen: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Data curation, Conceptualization. **Qingshan Hou:** Writing – review & editing, Validation, Methodology, Investigation. **Peng Cao:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Formal analysis, Conceptualization. **Jinzhu Yang:** Project administration, Funding acquisition, Formal analysis. **Huazhu Fu:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization. **Osmar R. Zaiane:** Writing – review & editing, Formal analysis, Conceptualization. **Zhaolin Chen:** Writing – review & editing, Supervision, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the Science and Technology Joint Project of Liaoning Province (2023JH2/101700367 and ZX20240193), the Fundamental Research Funds for the Central Universities (N25BJD005), and the National Natural Science Foundation of China (62076059).

References

[1] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–mICCAI 2015: 18th international conference, munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer; 2015, p. 234–41.

[2] Milletari F, Navab N, Ahmadi S-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). Ieee; 2016, p. 565–71.

[3] Xu Q, Zhang R, Fan Z, Wang Y, Wu Y-Y, Zhang Y. Fourier-based augmentation with applications to domain generalization. Pattern Recognit 2023;139:109474.

[4] Zhou K, Yang Y, Qiao Y, Xiang T. Mixstyle neural networks for domain generalization and adaptation. Int J Comput Vis 2024;132(3):822–36.

[5] Chen X, Shen Z, Cao P, Yang J, Zaiane OR. ConStyX: Content style augmentation for generalizable medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2025, p. 100–10.

[6] Yao H, Hu X, Li X. Enhancing pseudo label quality for semi-supervised domain-generalized medical image segmentation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, 2022, p. 3099–107.

[7] Zhang R, Xu Q, Huang C, Zhang Y, Wang Y. Semi-supervised domain generalization for medical image analysis. In: 2022 IEEE 19th international symposium on biomedical imaging. ISBI, IEEE; 2022, p. 1–5.

[8] Zhou K, Loy CC, Liu Z. Semi-supervised domain generalization with stochastic stylematch. Int J Comput Vis 2023;1–11.

[9] Sohn K, Berthelot D, Carlini N, Zhang Z, Zhang H, Raffel CA, Cubuk ED, Kurakin A, Li C-L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Adv Neural Inf Process Syst 2020;33:596–608.

[10] Yang Y, Soatto S. Fda: Fourier domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 4085–95.

[11] Xu Q, Zhang R, Zhang Y, Wang Y, Tian Q. A fourier-based framework for domain generalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, p. 14383–92.

[12] Zhou K, Yang Y, Qiao Y, Xiang T. Domain generalization with MixStyle. In: International conference on learning representations. 2021.

[13] Chen Z, Pan Y, Ye Y, Cui H, Xia Y. Treasure in distribution: a domain randomization based multi-source domain generalization for 2d medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2023, p. 89–99.

[14] Huang Y, Han L, Dou H. Generative feature style augmentation for domain generalization in medical image segmentation. Pattern Recognit 2025;162:111416.

[15] Xu Z, Liu D, Yang J, Raffel C, Niethammer M. Robust and generalizable visual representation learning via random convolutions. In: International conference on learning representations. 2021.

[16] Choi S, Das D, Choi S, Yang S, Park H, Yun S. Progressive random convolutions for single domain generalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 10312–22.

[17] Lyu F, Ye M, Carlsen JF, Erleben K, Darkner S, Yuen PC. Pseudo-label guided image synthesis for semi-supervised covid-19 pneumonia infection segmentation. IEEE Trans Med Imaging 2022;42(3):797–809.

[18] Mayr C, Kubler C, Haala N, Teutsch M. Narrowing the synthetic-to-real gap for thermal infrared semantic image segmentation using diffusion-based conditional image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024, p. 3131–41.

[19] Quinton F, Presles B, Popoff R, Godard F, Chevallier O, Pellegrinelli J, Vrigneaud J-M, Alberini J-L, Meriaudeau F. Tackling data scarcity: Synthetic tumour and mask generation to improve image segmentation. Artif Intell Med 2025;103348.

[20] Zhang L, Jindal B, Alaa A, Weinreb R, Wilson D, Segal E, Zou J, Xie P. Generative AI enables medical image segmentation in ultra low-data regimes. Nat Commun 2025;16(1):6486.

[21] Van Engelen JE, Hoos HH. A survey on semi-supervised learning. Mach Learn 2020;109(2):373–440.

[22] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Adv Neural Inf Process Syst 2017;30.

[23] Lee D-H, et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, 2013, p. 896.

[24] Berthelot D, Carlini N, Goodfellow I, Papernot N, Oliver A, Raffel CA. Mixmatch: A holistic approach to semi-supervised learning. Adv Neural Inf Process Syst 2019;32.

[25] Chen X, Yuan Y, Zeng G, Wang J. Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, p. 2613–22.

[26] Liu Y, Tian Y, Chen Y, Liu F, Belagiannis V, Carneiro G. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 4258–67.

[27] Yang L, Zhuo W, Qi L, Shi Y, Gao Y. St++: Make self-training work better for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 4268–77.

[28] Yang L, Qi L, Feng L, Zhang W, Shi Y. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 7236–46.

[29] Mai H, Sun R, Wu F. Relaxed class-consensus consistency for semi-supervised semantic segmentation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 39, 2025, p. 6045–53.

[30] Shen Z, Cao P, Yang H, Liu X, Yang J, Zaiane OR. Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation. In: Proceedings of the thirty-second international joint conference on artificial intelligence. IJCAI-23, 2023, p. 4199–207.

[31] Zhang Y, Jiao R, Liao Q, Li D, Zhang J. Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation. Artif Intell Med 2023;138:102476.

[32] Liu X, Li W, Yuan Y. Diffrect: Latent diffusion label rectification for semi-supervised medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2024, p. 56–66.

[33] Yang Y, Sun G, Zhang T, Wang R, Su J. Semi-supervised medical image segmentation via weak-to-strong perturbation consistency and edge-aware contrastive representation. Med Image Anal 2025;101:103450.

[34] Zeng Q, Lu Z, Xie Y, Xia Y. Pick: Predict and mask for semi-supervised medical image segmentation. Int J Comput Vis 2025;133(6):3296–311.

[35] Shen Z, Cao P, Su J, Yang J, Zaiane OR. Adaptive mix for semi-supervised medical image segmentation. Med Image Anal 2026;103857.

[36] Zeng Q, Luo H, Ma X, Lu Z, Hu Y, Lu M, Zhang Y, Xia Y. Decoupling target semantics via text-anchored visual contrast for semi-supervised medical image segmentation. IEEE Trans Image Process 2026.

[37] Zhou T, Gu Y, Huang K, Fu H, Wu X-j, Kittler J. Cross-sample consistency learning for semi-supervised medical image segmentation. Int J Comput Vis 2026;134(5):218.

[38] Castro DC, Walker I, Glocker B. Causality matters in medical imaging. Nat Commun 2020;11(1):3673.

[39] Li H, Li H, Zhao W, Fu H, Su X, Hu Y, Liu J. Frequency-mixed single-source domain generalization for medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2023, p. 127–36.

[40] Zhao H, Dong W, Yu R, Zhao Z, Du B, Xu Y. Morestyle: Relax low-frequency constraint of fourier-based image reconstruction in generalizable medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2024, p. 434–44.

[41] Lu J, Ma S, Liu Y, Ma Y, Mou L, Jiang Y, Zhao Y. Rethinking data augmentation for single-source domain generalization in OCT image segmentation. IEEE J Biomed Health Inform. 2025.

[42] Ouyang C, Chen C, Li S, Li Z, Qin C, Bai W, Rueckert D. Causality-inspired single-source domain generalization for medical image segmentation. IEEE Trans Med Imaging 2022;42(4):1095–106.

[43] Wang R, Guo J, Zhang J, Qi L, Yu Q, Shi Y. A hybrid dual-augmentation constraint framework for single-source domain generalization in medical image segmentation. Pattern Recognit 2026;170:112082.

[44] Nam J-H, Lee S-C. Frequency-based boundary-guided attention network for domain generalizable polyp segmentation from colonoscopy images. Artif Intell Med 2025;103288.

[45] Qi L, Yang H, Shi Y, Geng X. Multimatch: Multi-task learning for semi-supervised domain generalization. ACM Trans Multimed Comput Commun Appl 2024;20(6):1–21.

[46] Song J, Chen H, Qin J, Zhao N. Dual-supervised asymmetric co-training for semi-supervised medical domain generalization. IEEE Trans Multimed 2025.

[47] Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 6023–32.

[48] Hariharan B, Arbeláez P, Girshick R, Malik J. Hypercolumns for object segmentation and fine-grained localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, p. 447–56.

[49] Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision. ECCV, 2018, p. 801–18.

[50] Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip Rev: Comput Stat 2010;2(4):433–59.

[51] Candès EJ, Li X, Ma Y, Wright J. Robust principal component analysis? J ACM 2011;58(3):1–37.

[52] Gao S, Zhuang X. Rank-one network: An effective framework for image restoration. IEEE Trans Pattern Anal Mach Intell 2020;44(6):3224–38.

[53] Shen Z, Cao P, Yang J, Zaiane OR, Chen Z. Style content decomposition-based data augmentation for domain generalizable medical image segmentation. In: Medical imaging with deep learning. 2026.

[54] Zhuang X. Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE Trans Pattern Anal Mach Intell 2018;41(12):2933–46.

[55] Zhang Z, Yin FS, Liu J, Wong WK, Tan NM, Lee BH, Cheng J, Wong TY. Origa-light: An online retinal fundus image database for glaucoma analysis and research. In: 2010 annual international conference of the IEEE engineering in medicine and biology. IEEE; 2010, p. 3065–8.

[56] Sivaswamy J, Krishnadas S, Joshi GD, Jain M, Tabish AUS. Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In: 2014 IEEE 11th international symposium on biomedical imaging. ISBI, IEEE; 2014, p. 53–6.

[57] Orlando JI, Fu H, Breda JB, Van Keer K, Bathula DR, Diaz-Pinto A, Fang R, Heng P-A, Kim J, Lee J, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med Image Anal 2020;59:101570.