Learning Generalized Medical Image Segmentation from Decoupled Feature Queries

Qi Bi[1,2*], Jingjun Yi[1,2*], Hao Zheng[1†], Wei Ji[3], Yawen Huang[1], Yuexiang Li[4†], Yefeng Zheng[1]

[1]Jarvis Research Center, Tencent YouTu Lab, ShenZhen, China
[2]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China
[3]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada
[4]Medical AI ReSearch (MARS) Group, Guangxi Medical University, Nanning, China
{q_bi, rsjingjuny}@whu.edu.cn, yuexiangli@sr.gxmu.edu.cn, {howzheng, yefengzheng}@tencent.com

Abstract

Domain generalized medical image segmentation requires models to learn from multiple source domains and generalize well to arbitrary unseen target domain. Such a task is both technically challenging and clinically practical, due to the domain shift problem (i.e., images are collected from different hospitals and scanners). Existing methods focused on either learning shape-invariant representation or reaching consensus among the source domains. An ideal generalized representation is supposed to show similar pattern responses within the same channel for cross-domain images. However, to deal with the significant distribution discrepancy, the network tends to capture similar patterns by multiple channels, while different cross-domain patterns are also allowed to rest in the same channel. To address this issue, we propose to leverage channel-wise decoupled deep features as queries. With the aid of cross-attention mechanism, the long-range dependency between deep and shallow features can be fully mined via self-attention and then guides the learning of generalized representation. Besides, a relaxed deep whitening transformation is proposed to learn channel-wise decoupled features in a feasible way. The proposed decoupled feature query (DFQ) scheme can be seamlessly integrate into the Transformer segmentation model in an end-to-end manner. Extensive experiments show its state-of-the-art performance, notably outperforming the runner-up by 1.31% and 1.98% with DSC metric on generalized fundus and prostate benchmarks, respectively. Source code is available at https://github.com/BiQiWHU/DFQ.

Introduction

Despite the rapid development of deep learning techniques, most existing medical image segmentation approaches assume that the training and testing samples follow the same statistical distribution. Unfortunately, this assumption may not be fulfilled in many practical medical scenarios. In practice, it is notoriously taxing and expertise-demanding to annotate large amount of segmentation ground truth (Wang et al. 2020; Ouyang et al. 2020; Zhou, Qi, and Shi 2022; Cui et al. 2021; Yao, Hu, and Li 2022). In this regard, medical images are usually collected from a variety of hospitals and annotated by different annotators with different levels of expertise (Ji et al. 2021; Reiß et al. 2021). Consequently, the domain shift inevitably exists among these data sources, and thus leads to the high requirement on the generalization ability for medical image segmentation models.

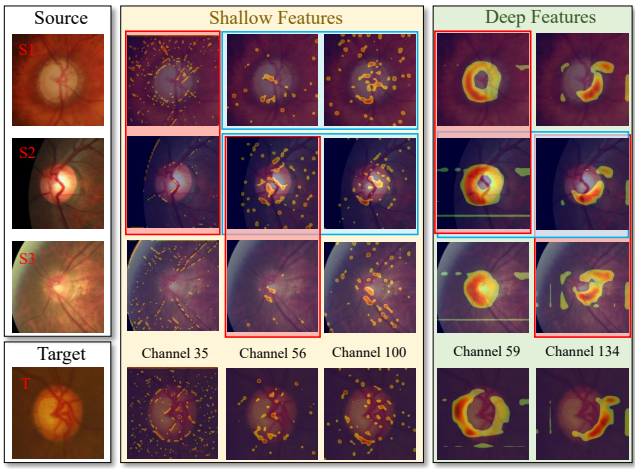

Figure 1: Key challenges to learn domain generalized medical segmentation (visualized with GradCAM). (1) Similar cross-domain features rest in multiple channels with redundancy (in blue boxes); (2) Cross-domain feature misalignment among the same channel (in red boxes).

In the past few years, the area of domain adaptation for medical image segmentation has been extensively studied. Its pre-requisite is that samples from the target domain are involved in training (Bian et al. 2021; You et al. 2022; Chen et al. 2019), and can only generalize to the target domain seen in training (Zhou, Qi, and Shi 2022). In contrast, the domain generalization paradigm allows the learnt representation to be generalized to any unseen target domains, which significantly alleviates the aforementioned dilemma of annotated data (Liu et al. 2021a; Wang et al. 2020; Zhou, Qi, and Shi 2022; Hu et al. 2023). In general, learning domain generalized medical image segmentation is both technically and clinically significant, as it predicts reliable segmentation results from a variety of scanners, annotators and hospitals.

Existing domain generalized medical image segmentation

*These authors contributed equally.

†Corresponding author

methods can be summarized into two categories. The one is to learn shape-invariant features from multiple source domains (Liu, Dou, and Heng 2020; Liu et al. 2021a), and the other one is to explicitly learn the inter-domain shift among multiple source domains (Wang et al. 2020; Zhou, Qi, and Shi 2022; Hu et al. 2023). Unfortunately, these methods may not be able to handle the feature distribution variation on arbitrary unseen domains under different imaging conditions (e.g., illumination, image contrast, and scanning).

Due to the aforementioned domain shift problem, medical images from different domains may have dramatically different activation patterns among the same channel of a deep learning model (shown in red boxes of Fig. 1). The feature misalignment is particularly obvious for shallower features, which are more sensitive to the variation of imaging conditions. To capture the required pattern in each domain, the network tends to learn a similar pattern in multiple channels, which further leads to varying degrees of feature redundancy across images from different domains (shown in blue boxes of Fig. 1). Feature redundancy helps the model to perform well on the training data from various domains even in the presence of single-channel mismatches, while negatively affecting the generalization ability to unseen domains.

In this paper, we are motivated to address the feature misalignment, which helps the models to build a more expressive cross-domain medical image representation and improves the their generalization on unseen target domains. First, we propose to minimize the channel-wise correlation among cross-domain medical images, which helps remove the feature redundancy and maximize the per-channel representation ability. A more expressive per-channel and less redundant representation in the feature encoding stage in turn benefits the generalization on arbitrary unseen target domains. To this end, we propose a relaxed deep whitening transformation, which can be integrated into existing deep segmentation models in a feasible and learnable fashion.

On the other hand, the feature de-correlation may not necessarily warrant medical images from different domains have similar activation patterns within the same channel. To further address the intra-channel feature misalignment, we turn to the decoding stage, and propose to use the self-attention mechanism as an implicit constriction. Specifically, we use the decoupled deeper features as the query, and the shallow features as the key and value. The inherent long-dependency between decoupled deeper and shallower features restricts the overall framework to learn domain generalized representation from the scratch. Overall, the proposed decoupled feature query (DFQ) learning scheme is integrated into Transformer segmentation models (Xie et al. 2021; Shim et al. 2023) in a learnable fashion.

Extensive experiments validate the effectiveness of the proposed DFQ on two standard domain generalized medical image segmentation benchmarks, namely, optic cup/disk segmentation on fundus images (Wang et al. 2020) and prostate segmentation on magnetic resonance imaging (MRI) (Liu, Dou, and Heng 2020). On both benchmarks, samples from one domain are used as unseen target domain, while samples from the rest domains are used as source domains. Finally, visualized segmentation predictions and feature space analysis are presented to further validate the effectiveness of the proposed method.

Our contributions can be summarized as follows.

- We propose to learn generalized medical image representation from decoupled feature queries (DFQ), which addresses the feature misalignment from cross-domain medical images. The proposed decoupled feature query scheme can be seamlessly integrated into Transformer segmentation models to achieve the better domain generalization performance.
- A relaxed deep whitening transformation is proposed to de-correlate the features in a learnable and flexible way.
- The proposed framework outperforms the state-of-the-art by at least 1.31% and 1.98% DSC on Fundus and Prostate benchmarks, respectively.

Related Work

Medical image segmentation has been developed rapidly owing to the stronger representation from deep learning techniques (Bi et al. 2022; Ji et al. 2022; Li et al. 2021a). In the early deep learning era, U-Net (Ronneberger, Fischer, and Brox 2015) and its variants (Zhou et al. 2018; Azad et al. 2021; Daza, Pérez, and Arbeláez 2021) were dominant for medical image segmentation. Later on, DeepLab (Chen et al. 2017) and its modifications (Gu et al. 2019; Feng et al. 2022) became the dominant trend. More recently, Vision Transformer (ViT) has shown stronger feature representation power than convolutional neural networks (Xie et al. 2021; Shim et al. 2023). Its self-attention mechanism is capable to mine the long-range dependencies (Liu et al. 2021b). Consequently, ViT based medical segmentation pipelines have recently drawn extensive attention (Cao et al. 2022; Gao, Zhou, and Metaxas 2021). In addition, medical image segmentation under weakly-supervised (Pan et al. 2022), semi-supervised (Wu et al. 2022) and multi-annotation (Ji et al. 2021) scenarios have also been studied.

Domain generalization has been extensively studied in both computer vision and machine learning communities under the non task-specific settings (Xu et al. 2021; Mahajan, Tople, and Sharma 2021; Li et al. 2021b). On the other hand, domain generalized segmentation in the computer vision community usually focuses on the driving scenes under the single domain generalization setting (Pan et al. 2018; Huang et al. 2019; Peng et al. 2022; Pan et al. 2019; Choi et al. 2021; Xu et al. 2022; Peng et al. 2022; Lee et al. 2022; Zhao et al. 2022; Zhong et al. 2022; Li et al. 2023; Bi, You, and Gevers 2023). In contrast, the key challenges in generalized medical image segmentation lie in the great style variations from multiple source domains for training.

Domain generalized medical image segmentation intends to learn a semantic representation generalized to any unseen target domain by learning from only source domains. Specially, (Liu, Dou, and Heng 2020) incorporated meta-learning to learn shape robustness representation. (Zhang et al. 2020) proposed a deep staked transformation, which augmented the images from all domains under a variety of

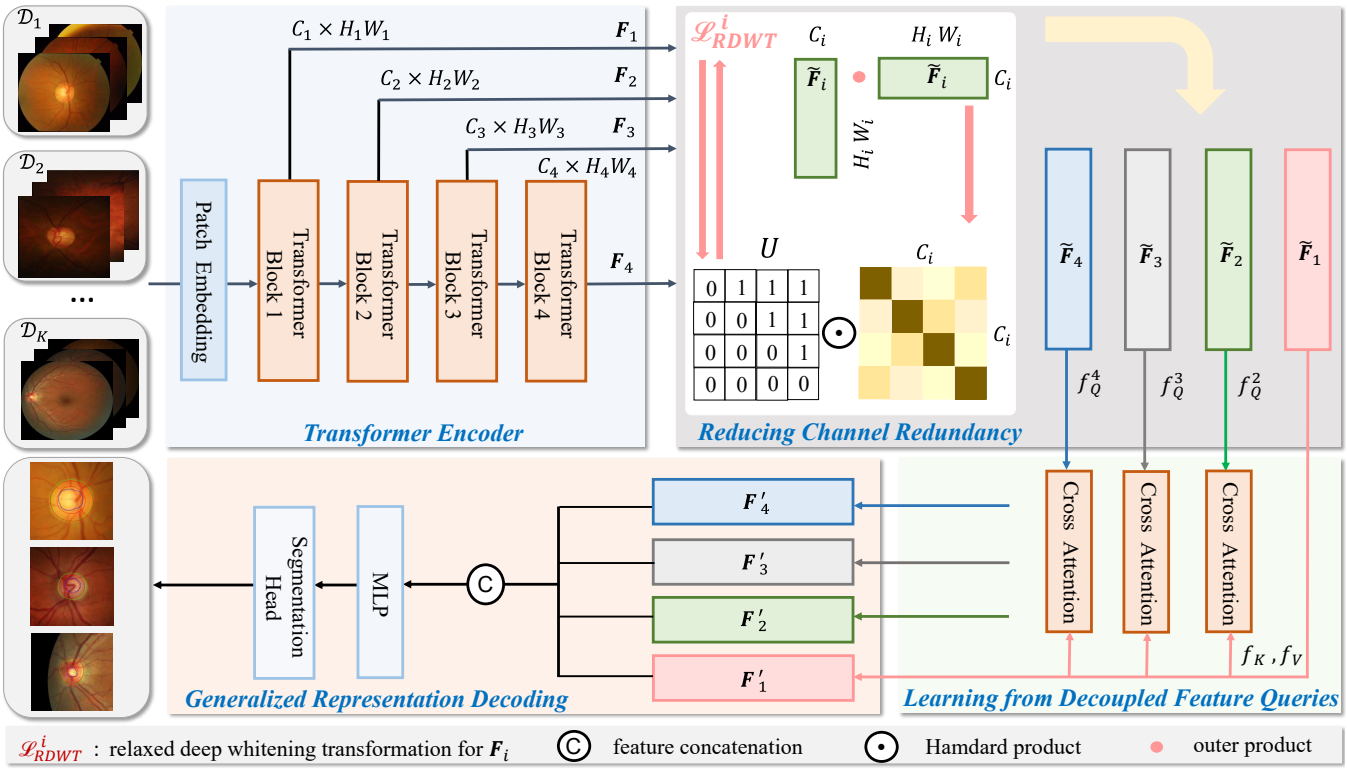

Figure 2: Framework of the proposed Decoupled Feature Query (DFQ). After feature extraction from a Transformer encoder, it consists three key steps, namely, reducing channel redundancy, learning from decoupled feature queries, and generalized representation learning. Reducing channel redundancy is implemented by our proposed relaxed deep whitening transformation.

data augmentation transformations. (Liu et al. 2021a) proposed a boundary-oriented episodic learning to enhance the shape robustness. (Wang et al. 2020) focused on learning robust medical semantics under the style variations. (Zhou, Qi, and Shi 2022) used the mixup strategy to enhance the shape diversity, and an additional reconstruction head to learn the domain diversity. (Hu et al. 2023) focused on the content enhancement for generalized medical image segmentation.

Methodology

Problem Formulation & Overview

Given K source domains $\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_K$ and an unseen target domain $\mathcal{D}_{K+1}$. For domain $\mathcal{D}_k$, the joint image and label pair is denoted as $\{(x_n^{(k)}, y_n^{(k)})\}_{n=1}^{N_k}$, where $k = 1, 2, \cdots, K$, and N_k refers to the sample number in domain $\mathcal{D}_k$. The learning objective is to learn a segmentation model $F_\theta : x \rightarrow y$ using all the source domains $\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_K$, and generalize well on the unseen target domain $\mathcal{D}_{K+1} = \{(x_n^{(K+1)})\}_{n=1}^{N_{K+1}}$.

Fig. 2 gives an overview of the proposed framework, where $\mathbf{F}_1 \in \mathbb{R}^{C_1 \times H_1 W_1}$, $\mathbf{F}_2 \in \mathbb{R}^{C_2 \times H_2 W_2}$, $\mathbf{F}_3 \in \mathbb{R}^{C_3 \times H_3 W_3}$ and $\mathbf{F}_4 \in \mathbb{R}^{C_4 \times H_4 W_4}$ denote the image features from the first, second, third and forth Transformer block (from high to low resolution), respectively. After feature encoding of the medical image, three key steps (namely reducing channel redundancy, learning from de-correlated feature queries, and decoding generalized representation) are involved to yield the robust feature for domain generalized medical image segmentation.

Reducing Channel Redundancy

To deal with the distribution discrepancy across domains, deep neural network tends to extract similar patterns in multiple channels, which inevitably leads to feature redundancy. The decoupling of channel-wise correlation can reduce the redundancy and maximize per-channel representation ability, which helps to learn more expressive features generalizable to arbitrary source domains and unseen target domains.

A common channel de-correlation approach is the whitening transformation (Li et al. 2017). Take the image feature from the first block of the Transformer encoder $\mathbf{F}_1 \in \mathbb{R}^{C_1 \times H_1 W_1}$ as an example. Its whitened feature $\widetilde{\mathbf{F}}_1$ can be mathematically computed as

$$\widetilde{\mathbf{F}}_1 = \mathbf{\Sigma}_\mu^{-\frac{1}{2}}(\mathbf{F}_1 - \boldsymbol{\mu} \cdot \mathbf{1}^{\mathrm{T}}), \tag{1}$$

where the mean vector and covariance matrix can be computed as

$$\boldsymbol{\mu} = \frac{1}{HW}\mathbf{F}_1 \cdot \mathbf{1} \in \mathbb{R}^{C_1 \times 1}, \tag{2}$$

$$\mathbf{\Sigma}_\mu = \frac{1}{HW}(\mathbf{F}_1 - \boldsymbol{\mu} \cdot \mathbf{1}^{\mathrm{T}})(\mathbf{F}_1 - \boldsymbol{\mu} \cdot \mathbf{1}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{C_1 \times C_1}. \tag{3}$$

To integrate the whitening transformation in a learnable fashion, a basic solution is to follow (Cho et al. 2019). The

so-called *deep whitening transformation* (DWT) drives $\mathbf{\Sigma}_\mu$ towards the identity matrix $\mathbf{I} \in \mathbb{R}^{C_1 \times C_1}$,

$$\mathcal{L}_{DWT}^{\mathbf{F}_1} = \mathbb{E}[||\mathbf{\Sigma}_\mu - \mathbf{I}||_1]. \tag{4}$$

Let $\mathbf{\Sigma}_\mu(i,i)$ and $\mathbf{\Sigma}_\mu(i,j)$ denote a diagonal and an off-diagonal element of $\mathbf{\Sigma}_\mu$ respectively, where $0 \leq i,j \leq N$, and $i \neq j$. Then, denoting $\mathbf{F}_1^\dagger = \mathbf{F}_1 - \boldsymbol{\mu} \cdot \mathbf{1}^{\mathrm{T}}$, Eq. 4 can be decomposed as

$$||\mathbf{\Sigma}(i,i) - 1||_1 = ||\frac{|\mathbf{F}_{1,i}^\dagger||\mathbf{F}_{1,i}^\dagger|}{W \cdot H} - 1||_1, \tag{5}$$

$$||\mathbf{\Sigma}(i,j)||_1 = ||\frac{|\mathbf{F}_{1,i}^\dagger||\mathbf{F}_{1,j}^\dagger| \cos\theta}{W \cdot H}||_1. \tag{6}$$

Eq. 5 poses a numerical constraint on the diagonal, and ideally, the impact of Eq. 6 in the feature space is to decouple $\mathbf{F}_{1,i}^\dagger$ and $\mathbf{F}_{1,j}^\dagger$ by forcing the off-diagonal to be orthogonal. However, $||\mathbf{\Sigma}(i,j)||_1$ can also be reduced by decreased $|\mathbf{F}_{1,i}^\dagger|$ or $|\mathbf{F}_{1,j}^\dagger|$, providing a shortcut for reducing $\mathcal{L}_{WT}^{\mathbf{F}_1}$ when learning decorrelated representations is extremely harder for some channels. Consequently, it can be found that the channel correlation still exists in the results of DWT.

To resolve this problem, we normalize the $\mathbf{F}_1$ by

$$\widetilde{\mathbf{F}}_1 = \frac{\mathbf{F}_1 - \boldsymbol{\mu} \cdot \mathbf{1}^{\mathrm{T}}}{\boldsymbol{\sigma} \cdot \mathbf{1}^{\mathrm{T}}} \tag{7}$$

before calculating the covariance matrix. After that, $|\widetilde{\mathbf{F}}_{1,i}| = 1$, $||\mathbf{\Sigma}(i,i) - 1||_1$ becomes a constant and $||\mathbf{\Sigma}(i,j)||_1$ is only correlated with the angle between two channels. We can use a strict upper triangular matrix $\mathbf{U}$ to approximate the learning objective,

$$\mathbf{U}_{(i,j)} = \begin{cases} 1 & i<j \\ 0 & i \geq j \end{cases} \quad 0 \leq i,j \leq N, \tag{8}$$

$$\mathcal{L}_{RDWT}^{\mathbf{F}_1} = \mathbb{E}[||\mathbf{\Sigma}_\mu \odot \mathbf{U}||_1], \tag{9}$$

where $\odot$ denotes the Hadamard product. Compared with the original DWT, the magnitude constraint on the diagonal is relaxed by the normalized input. $\mathcal{L}_{RDWT}^{\mathbf{F}_1}$ only focuses on the correlation between channels and is more effective to reduce the feature redundancy. Moreover, under the supervision of $\mathcal{L}_{RDWT}^{\mathbf{F}_1}$, $\widetilde{\mathbf{F}}_1$ can be directly used as the whitening-transformed feature in the following.

For $\mathbf{F}_2$, $\mathbf{F}_3$ and $\mathbf{F}_4$, similarly, we can learn their whitening transformation by minimizing $\mathcal{L}_{RDWT}^{\mathbf{F}_2}$, $\mathcal{L}_{RDWT}^{\mathbf{F}_3}$ and $\mathcal{L}_{RDWT}^{\mathbf{F}_4}$, respectively. For simplicity, their whitened counterpart is denoted as $\widetilde{\mathbf{F}}_2$, $\widetilde{\mathbf{F}}_3$ and $\widetilde{\mathbf{F}}_4$, respectively.

Learning from De-correlated Feature Queries

The channel-wise decoupled features enhance the representation ability of deep neural networks in cross-domain scenarios. However, the relaxed whitening transformation loss cannot warrant that medical images from different domains show similar per-channel feature response, which is also crucial for learning a generalized model to unseen target datasets. To this end, we turn to the long-dependency inherent in the self-attention mechanism. Compared with deep semantic features, the shallow features directly face the distribution discrepancy across domains, resulting in severer intra-channel misalignment. When decoding the high-level features, the query is generated from deep features while the key and value are based on shallow features. The feature misalignment in shallow layers can lead to different attention maps in such a self-attention process, which further results in unstable feature aggregation for the decoding of deep representations. Under this correlation, the deep feature queries impose an implicit constraint on the consistency of shallow representations across different domains.

Specifically, for the channel-wise decoupled features $\widetilde{\mathbf{F}}_i$ from the i^{th} Transformer block, where $i = 2,3,4$, a linear transformation f_Q^i is used to generate the query,

$$\mathbf{Q}^i = f_Q^i(\widetilde{\mathbf{F}}_i). \tag{10}$$

For the channel-wise decoupled features from the first Transformer block, the key and value can be computed as

$$\mathbf{K} = f_K(\widetilde{\mathbf{F}}_1), \mathbf{V} = f_V(\widetilde{\mathbf{F}}_1), \tag{11}$$

where f_K and f_V are the linear transformations to generate the key and value, respectively.

Then, the cross-attention for the features from the i^{th} Transformer block can be computed as

$$\mathrm{Attention}(\mathbf{Q}^i, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}(\frac{\mathbf{Q}^i\mathbf{K}}{\sqrt{d_k}})\mathbf{V}, \tag{12}$$

where $\mathrm{Softmax}$ denotes the softmax normalization function.

After the feed-forward layer and normalization, let $\mathbf{F}_1^{'}, \mathbf{F}_2^{'}, \mathbf{F}_3^{'}$ and $\mathbf{F}_4^{'}$ denote the learnt generalized representation from the first, second, third and forth Transformer blocks, respectively.

Decoding Generalized Representation

The final step is to decode these generalized representations for medical segmentation prediction. This feature fusion is implemented by a linear layer parameterized by weight $\mathbf{W}_1$ and bias $\mathbf{b}_1$, presented as

$$\mathbf{F} = \mathbf{W}_1[\mathbf{F}_1^{'}, \mathbf{F}_2^{'}, \mathbf{F}_3^{'}, \mathbf{F}_4^{'}] + \mathbf{b}_1, \tag{13}$$

where $[\cdot,\cdot]$ denotes the concatenation function.

Then, $\mathbf{F}$ is fed into the semantic segmentation head for final prediction. The standard binary cross-entropy loss and Dice loss, which for simplicity we denote as $\mathcal{L}_{seg}$, are used to minimize the difference between the final prediction and the ground truth.

The total loss function $\mathcal{L}$ is a combination between $\mathcal{L}_{seg}$ and $\mathcal{L}_{RDWT}^i$ for each feature,

$$\mathcal{L} = \mathcal{L}_{seg} + \lambda \cdot \sum_{i=1}^{4} \mathcal{L}_{RDWT}^i, \tag{14}$$

where we set λ to be 1×10^{-4}. Notice that, $\mathcal{L}_{RDWT}^i$ is a channel-wise sum while $\mathcal{L}_{seg}$ is sample-wise and single-channel.

Method	Domain-1		Domain-2		Domain-3		Domain-4		Domain-5		Domain-6		Average	
	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓
Intra-domain	89.53	1.39	88.42	1.44	87.65	1.67	83.01	3.58	83.39	2.99	84.97	2.00	86.16	2.18
DeepAll	89.16	2.09	87.31	1.27	74.12	3.02	88.85	2.36	83.22	3.51	88.39	1.67	85.18	2.32
BigAug	90.68	1.80	89.52	1.00	84.86	1.86	89.04	1.59	73.24	5.94	89.10	1.16	86.07	2.23
SAML	91.00	1.26	89.26	1.12	85.76	1.87	89.60	1.21	81.60	3.29	89.91	0.96	87.86	1.62
FedDG	91.41	1.29	89.95	0.97	85.10	2.63	89.13	1.51	76.69	4.52	90.63	1.03	87.15	1.99
DoFE	89.79	1.33	87.42	1.57	84.90	2.13	88.56	1.52	86.47	1.93	87.72	1.33	87.48	1.64
RAM-DSIR	87.56	1.04	90.20	0.81	86.92	2.23	88.72	1.16	87.17	1.81	87.93	1.15	88.08	1.37
DCAC	**91.76**	0.98	90.51	0.89	86.30	**1.77**	89.13	1.53	83.39	2.46	90.56	0.85	88.61	1.41
DFQ (Ours)	88.28	**0.84**	**91.66**	**0.63**	**89.00**	2.24	**90.16**	**0.67**	**89.57**	**1.43**	**90.83**	**0.59**	**89.92**	**1.07**

Table 1: Performance comparison of the proposed method and existing methods on domain generalized prostate segmentation.

Method	Domain-1		Domain-2		Domain-3		Domain-4		Average	
	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓
Intra-domain	80.06	20.13	73.13	24.91	83.80	11.20	84.46	8.99	86.46	13.58
DeepAll	79.04	20.32	73.02	24.99	82.26	12.01	84.85	8.39	85.75	13.91
BigAug	80.37	19.50	74.73	22.64	85.39	10.07	86.47	8.32	86.88	13.25
SAML	81.03	19.31	76.61	19.31	85.40	9.99	86.06	8.86	87.60	12.36
FedDG	81.66	18.79	76.31	19.98	85.23	10.86	85.27	8.94	87.29	12.64
DoFE	81.95	18.59	78.31	16.40	85.51	10.06	86.61	8.28	88.14	11.61
RAM-DSIR	85.48	16.05	78.82	14.01	87.44	9.02	85.84	8.29	88.94	10.32
DCAC	81.43	19.20	77.72	17.15	86.80	9.14	**87.68**	7.12	88.47	11.32
DFQ (Ours)	**87.30**	**15.72**	**81.92**	**13.05**	**88.95**	**7.70**	87.47	**6.55**	**90.57**	**9.52**

Table 2: Performance comparison of the proposed method and existing methods on domain generalized optic cup segmentation. Average refers to the average of optic & disk results on all four settings. Best performance is highlighted in bold.

Method	Domain-1		Domain-2		Domain-3		Domain-4	
	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓
Intra-domain	95.82	7.53	87.79	18.75	93.20	9.64	93.41	7.51
DeepAll	95.82	7.63	87.34	18.70	91.37	11.40	92.27	7.83
BigAug	95.59	7.75	87.40	18.89	92.04	11.09	93.05	7.75
SAML	95.74	7.66	87.29	19.20	93.92	8.62	94.76	5.90
FedDG	95.47	7.81	86.34	19.57	93.36	9.12	94.68	6.02
DoFE	96.04	7.05	89.20	15.75	93.23	9.76	94.28	6.99
RAM-DSIR	95.75	7.12	89.43	13.86	94.67	7.11	94.10	7.06
DCAC	**96.54**	6.35	87.85	18.28	94.28	8.11	**95.40**	**5.20**
DFQ (Ours)	6.50	**6.01**	**92.52**	**12.09**	**95.04**	**7.05**	94.85	5.84

Table 3: Performance comparison of the proposed method and existing methods on domain generalized optic disk segmentation. Best performance is highlighted in bold.

Implementation Details

Mix Transformer (MiT-B3) (Xie et al. 2021) is used as the backbone. For the final MLP before the segmentation head, the embedding dimension is set 768. Following prior work (Zhou, Qi, and Shi 2022), the model was trained 400 epochs with an initial learning rate 5×10^{-4} on the Fundus benchmark, and 200 epochs with an initial learning rate 3×10^{-4} on the Prostate benchmark.

The data pre-processing strictly follows the prior works (Wang et al. 2020; Zhou, Qi, and Shi 2022), where the fundus images were firstly centered cropped into a size of 800×800 pixels. Both prostate images and the cropped fundus images were resized into 256×256 pixels as input.

Experiments & Analysis

Datasets & Evaluation Protocols

DG Fundus benchmark (Wang et al. 2020) consists of four optic cup/disc segmentation datasets, namely, Drishti-GS (Sivaswamy et al. 2015), RIM-ONE-r3 (Fumero et al. 2011), REFUGE (train) (Orlando et al. 2020), and REFUGE (val) (Orlando et al. 2020), which we denote as Domain-1, Domain-2, Domain-3 and Domain-4, respectively.

DG Prostate benchmark (Liu, Dou, and Heng 2020) consists of 116 T2-weighted MRI cases from six domains, which we denote from Domain-1 to Domain-6.

Evaluation metrics include Dice Similarity Coefficient (DSC) and Average Surface Distance (ASD), which strictly follow the prior medical image segmentation works.

Comparison with SOTA

The compared state-of-the-art domain generalized medical segmentation methods include BigAug (Zhang et al. 2020), SAML (Liu, Dou, and Heng 2020), FedDG (Liu et al. 2021a), DoFE (Wang et al. 2020), RAM-DSIR (Zhou, Qi, and Shi 2022) and DCAC (Hu et al. 2023). Following prior works, two baseline settings are involved: Intra-domain refers to training and testing on the same domain, while DeepAll refers to aggregating samples from all source do-

mains for training a deep model.

Results on the Prostate benchmark are reported in Table 1. The proposed method significantly outperforms existing state-of-the-art methods. Compared with the second-best, the ASD metric on the first, second, forth, fifth and sixth domain is improved by 0.14%, 0.18%, 0.49%, 0.38% and 0.26%, respectively. The DSC metric also outperforms all existing methods on five out of six domains by up to 2.08%. Besides, we achieve the state-of-the-art average DSC of 89.92% and average ASD of 1.07%. Compared with the recent state-of-the-art RAM-DSIR (Zhou, Qi, and Shi 2022) and DCAC (Hu et al. 2023), the average ASD is improved by 0.30% and 0.34%, respectively.

Results on the Fundus benchmark are reported in Table 2 and Table 3. The proposed method significantly outperforms existing state-of-the-art methods. Compared with the second-best RAM-DSIR, it achieves an average DSC gain of 1.63% and ASD improvement of 0.80%. On all the four target domains, the ASD of optic cup outperforms the second-best by up to 1.52%. On three of four target domains, the ASC of optic cup outperforms the state-of-the-art by up to 1.77%. On three out of four target domains, the DSC of optic cup outperforms the state-of-the-art by up to 3.10%.

Ablation Studies

On Each Component. The proposed DFQ framework consists of three key components, namely, segmentation backbone, feature as query and relaxed deep whitening transformation (RDWT), which we denote as Bb., FQ and RT, respectively. For fair evaluation, when there is no FQ component, the features from the backbone are directly fused into the segmentation head by an MLP. Table 4 reports the results on the Fundus benchmark. The use of feature queries leads to an improvement of 0.94% in DSC and 0.88% in ASD against the baseline. Our RT can further leads to an improvement of 1.16% in DSC and 0.68% in ASD.

On Each Scale. The decoupled key and value are from the first Transformer block, which we denote as $\mathbf{F}_1$. The decoupled queries are from the second, third and forth Transformer blocks, which we denote as $\mathbf{F}_2$, $\mathbf{F}_3$ and $\mathbf{F}_4$, respectively. We investigate the impact from decoupled queries and style-invariant key & values. Results are reported in Table 5. Both using style decoupled key and value ($\mathbf{F}_1$) and using style decoupled queries ($\mathbf{F}_2$, $\mathbf{F}_3$, $\mathbf{F}_4$) positively contribute to the segmentation results, but the use of style decoupled queries contribute more to the overall performance. The style-decoupling queries on $\mathbf{F}_2$, $\mathbf{F}_3$, $\mathbf{F}_4$ lead to a DSC improvement of 1.33%, 1.14%, 0.85%, and an ASD improvement of 0.19%, 0.10% and 0.12%, respectively.

Understanding DFQ

To evaluate if the channel redundancy and misalignment is well handled by the proposed DFQ, we compare it with the baseline (with only a Transformer encoder, feature query and a decoder) and DFQ with conventional deep whitening transformation. We denote the three settings as RDWT, Baseline and DWT, respectively.

On Reducing Channel Redundancy. The feature queries from the first to the forth block are computed with the covariance matrix, and are visualized in Fig. 3. The more yellow/purple, the higher/lower response. Experiments are conducted on the Prostate benchmark. From the left to right, we show the covariance matrix from the baseline, DWT, and RDWT, respectively. Ideally, a fully channel-decoupled covariance matrix eliminates the responses from all the off-diagonal regions. The proposed DFQ scheme shows the best

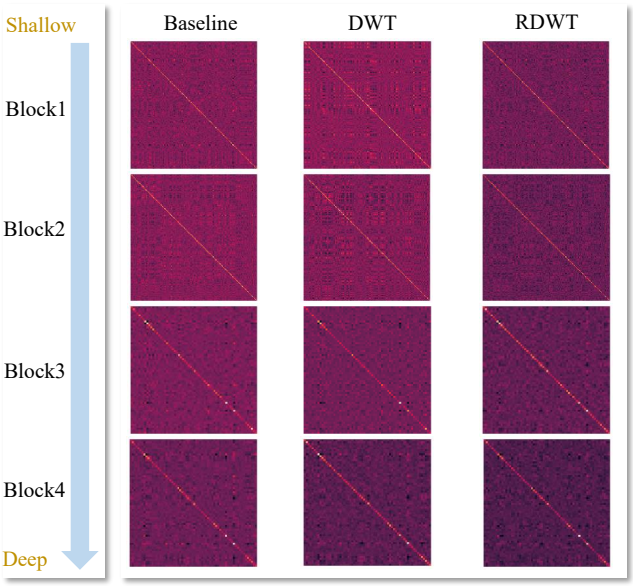

Figure 3: Visualization of the covariance matrix of feature queries, extracted by: baseline (left), conventional deep whitening transformation (DWT; middle), and the proposed relaxed deep whitening transformation (RDWT; right). The more yellow/purple, the higher/lower response.

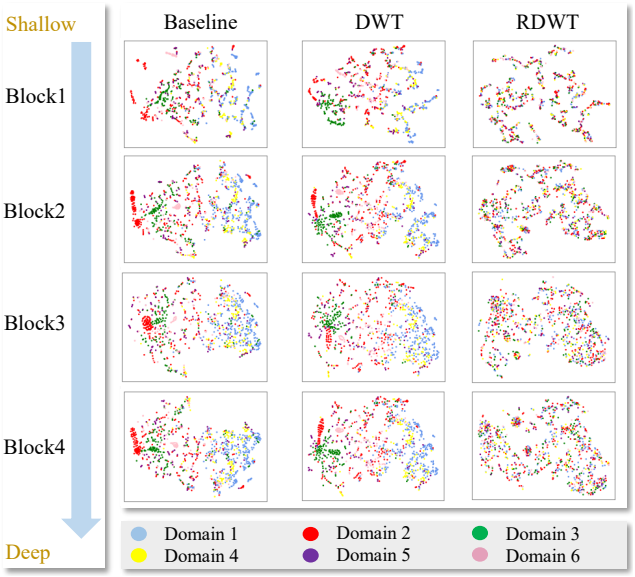

Figure 4: T-SNE visualization of the feature space from baseline (left), conventional deep whitening transformation (DWT; middle), and the proposed relaxed deep whitening transformation (RDWT; right). Zoom in for a better view.

Component			Domain-1		Domain-2		Domain-3		Domain-4	
Bb.	FQ	RT	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓
✓			84.92	17.29	78.49	15.94	86.28	10.30	85.54	8.35
✓	✓		86.16	16.48	80.39	13.54	87.13	9.68	86.69	7.45
✓	✓	✓	**87.30**	**15.72**	**81.92**	**13.05**	**88.95**	**7.70**	**87.47**	**6.55**

Table 4: Ablation studies on each component. Experiments on the optic cup segmentation of the Fundus benchmark.

Style-decoupling				Domain-1		Domain-2		Domain-3		Domain-4		Domain-5		Domain-6	
$\mathbf{F}_1$	$\mathbf{F}_2$	$\mathbf{F}_3$	$\mathbf{F}_4$	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓
✓				85.87	1.25	88.15	0.86	84.55	2.99	87.92	0.87	85.83	2.05	87.29	0.86
✓	✓			87.41	0.99	89.40	0.81	85.97	2.56	88.55	0.92	87.56	1.72	88.74	0.75
✓	✓	✓		87.82	0.92	90.97	0.68	87.89	2.48	89.86	0.75	88.35	1.56	89.58	0.73
✓	✓	✓	✓	**88.28**	**0.84**	**91.66**	**0.63**	**89.00**	**2.24**	**90.16**	**0.67**	**89.57**	**1.43**	**90.83**	**0.59**

Table 5: Ablation Studies on the style-decoupled queries from different blocks. Experiments on the Prostate benchmark.

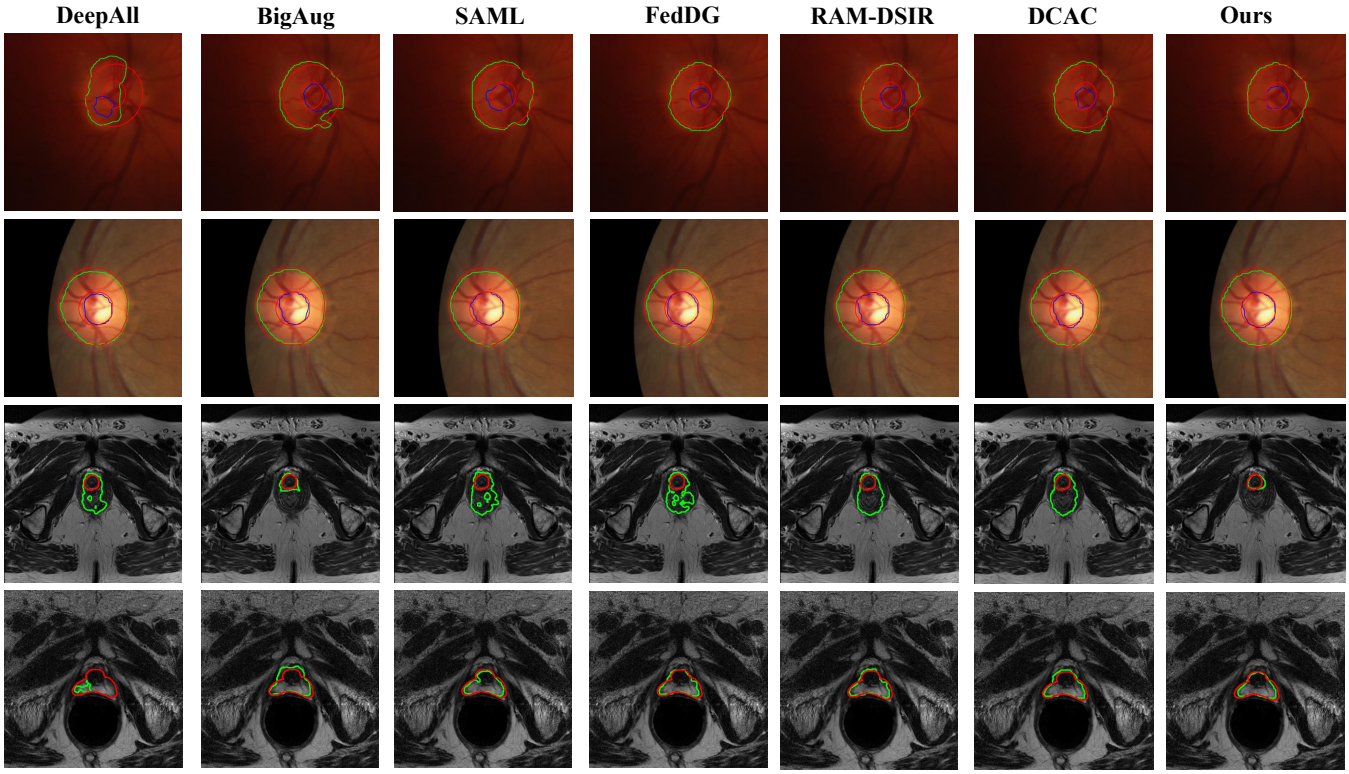

Figure 5: Exemplar domain generalized segmentation results of the proposed method and the state-of-the-art methods. The first and second rows are results from the Fundus benchmark. The third and forth rows are results from the Prostate benchmark. Ideally, the green and blue segmentation predictions should coincide the red ground truth. Zoom in for a better view.

performance in eliminating the off-diagonal elements.

On Cross-domain Feature Alignment. From left to right, Fig. 4 shows the feature space of the baseline, DWT and RDWT by t-SNE visualization. The proposed DFQ allows the samples from different domains to be more uniformly mixed, and thus helps minimize the domain gap.

Visualized Segmentation Results

Some exemplar segmentation results are visualized in Fig. 5. Compared with existing methods, the proposed method shows a more precise and reasonable prediction.

Conclusion

In this paper, we proposed a decoupled feature query (DFQ) learning scheme for domain generalized medical image segmentation, which aims to address the feature misalignment among cross-domain medical images. To enhance the per-channel representation ability and reduce the channel redundancy, we proposed a relaxed deep whitening transformation (RDWT). To learn similar channel-wise feature patterns from different domains, we innovatively used decoupled deep features as queries to guide the entire framework. Extensive experiments show the state-of-the-art performance of the proposed method.

References

Azad, R.; Bozorgpour, A.; Asadi-Aghbolaghi, M.; Merhof, D.; and Escalera, S. 2021. Deep frequency re-calibration U-Net for medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3274–3283.

Bi, Q.; You, S.; and Gevers, T. 2023. Learning Content-enhanced Mask Transformer for Domain Generalized Urban-Scene Segmentation. *arXiv preprint arXiv:2307.00371*.

Bi, Q.; Zhou, B.; Qin, K.; Ye, Q.; and Xia, G.-S. 2022. All grains, one scheme (AGOS): Learning multigrain instance representation for aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–17.

Bian, C.; Yuan, C.; Ma, K.; Yu, S.; Wei, D.; and Zheng, Y. 2021. Domain adaptation meets zero-shot learning: an annotation-efficient approach to multi-modality medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(5): 1043–1056.

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2022. Swin-Unet: Unet-like pure Transformer for medical image segmentation. In *European Conference on Computer Vision*, 205–218.

Chen, C.; Dou, Q.; Chen, H.; Qin, J.; and Heng, P.-A. 2019. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 865–872.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Cho, W.; Choi, S.; Park, D. K.; Shin, I.; and Choo, J. 2019. Image-to-image translation via group-wise deep whitening-and-coloring transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10639–10647.

Choi, S.; Jung, S.; Yun, H.; Kim, J. T.; Kim, S.; and Choo, J. 2021. RobustNet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11580–11590.

Cui, H.; Wei, D.; Ma, K.; Gu, S.; and Zheng, Y. 2021. A unified framework for Generalized low-shot medical image segmentation with scarce data. *IEEE Transactions on Medical Imaging*, 40(10): 2656–2671.

Daza, L.; Pérez, J. C.; and Arbeláez, P. 2021. Towards robust general medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention*, 3–13.

Feng, W.; Wang, L.; Ju, L.; Zhao, X.; Wang, X.; Shi, X.; and Ge, Z. 2022. Unsupervised domain adaptive fundus image segmentation with category-level regularization. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 497–506.

Fumero, F.; Alayón, S.; Sanchez, J. L.; Sigut, J.; and Gonzalez-Hernandez, M. 2011. RIM-ONE: An open retinal image database for optic nerve evaluation. In *International Symposium on Computer-based Medical Systems*, 1–6.

Gao, Y.; Zhou, M.; and Metaxas, D. N. 2021. UTNet: a hybrid Transformer architecture for medical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 61–71.

Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; and Liu, J. 2019. CE-Net: Context encoder network for 2D medical image segmentation. *IEEE Transactions on Medical Imaging*, 38(10): 2281–2292.

Hu, S.; Liao, Z.; Zhang, J.; and Xia, Y. 2023. Domain and content adaptive convolution based multi-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(1): 233–244.

Huang, L.; Zhou, Y.; Zhu, F.; Liu, L.; and Shao, L. 2019. Iterative normalization: beyond standardization towards efficient whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4874–4883.

Ji, W.; Li, J.; Bi, Q.; Liu, J.; Cheng, L.; et al. 2022. Promoting Saliency From Depth: Deep Unsupervised RGB-D Saliency Detection. In *International Conference on Learning Representations*.

Ji, W.; Yu, S.; Wu, J.; Ma, K.; Bian, C.; Bi, Q.; Li, J.; Liu, H.; Cheng, L.; and Zheng, Y. 2021. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12341–12351.

Lee, S.; Seong, H.; Lee, S.; and Kim, E. 2022. WildNet: Learning domain generalized semantic segmentation from the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9936–9946.

Li, J.; Ji, W.; Bi, Q.; Yan, C.; Zhang, M.; Piao, Y.; Lu, H.; et al. 2021a. Joint semantic mining for weakly supervised RGB-D salient object detection. *Advances in Neural Information Processing Systems*, 34: 11945–11959.

Li, L.; Gao, K.; Cao, J.; Huang, Z.; Weng, Y.; Mi, X.; Yu, Z.; Li, X.; and Xia, B. 2021b. Progressive domain expansion network for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 224–233.

Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; and Yang, M.-H. 2017. Universal style transfer via feature transforms. *Advances in Neural Information Processing Systems*, 30.

Li, Y.; Zhang, D.; Keuper, M.; and Khoreva, A. 2023. Intra-source style augmentation for improved domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 509–519.

Liu, Q.; Chen, C.; Qin, J.; Dou, Q.; and Heng, P.-A. 2021a. 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*, 1013–1023.

Liu, Q.; Dou, Q.; and 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*, 475–485.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin Transformer: Hierarchical vision Transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Mahajan, D.; Tople, S.; and Sharma, A. 2021. Domain generalization using causal matching. In *International Conference on Machine Learning*, 7313–7324.

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. *Medical Image Analysis*, 59: 101570.

Ouyang, C.; Biffi, C.; Chen, C.; Kart, T.; Qiu, H.; and Rueckert, D. 2020. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *European Conference Computer Vision*, 762–780.

Pan, J.; Bi, Q.; Yang, Y.; Zhu, P.; and Bian, C. 2022. Label-efficient hybrid-supervised learning for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2026–2034.

Pan, X.; Luo, P.; Shi, J.; and Tang, X. 2018. Two at once: Enhancing learning and generalization capacities via IBN-Net. In *European Conference on Computer Vision*, 464–479.

Pan, X.; Zhan, X.; Shi, J.; Tang, X.; and Luo, P. 2019. Switchable whitening for deep representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1863–1871.

Peng, D.; Lei, Y.; Hayat, M.; Guo, Y.; and Li, W. 2022. Semantic-aware domain generalized segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2594–2605.

Reiß, S.; Seibold, C.; Freytag, A.; Rodner, E.; and Stiefelhagen, R. 2021. Every annotation counts: Multi-label deep supervision for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9532–9542.

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

Shim, J.-h.; Yu, H.; Kong, K.; and Kang, S.-J. 2023. FeedFormer: Revisiting Transformer decoder for efficient semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2263–2271.

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 Biomedical Imaging Data Papers*, 2(1): 1004.

Wang, S.; Yu, L.; Li, K.; Yang, X.; Fu, C.-W.; and Heng, P.-A. 2020. Dofe: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. *IEEE Transactions on Medical Imaging*, 39(12): 4237–4248.

Wu, Y.; Ge, Z.; Zhang, D.; Xu, M.; Zhang, L.; Xia, Y.; and Cai, J. 2022. Mutual consistency learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 81: 102530.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with Transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Xu, Q.; Yao, L.; Jiang, Z.; Jiang, G.; Chu, W.; Han, W.; Zhang, W.; Wang, C.; and Tai, Y. 2022. DIRL: Domain-invariant representation learning for generalizable semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2884–2892.

Xu, Q.; Zhang, R.; Zhang, Y.; Wang, Y.; and Tian, Q. 2021. A Fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14383–14392.

Yao, H.; Hu, X.; and Li, X. 2022. Enhancing pseudo label quality for semi-supervised domain-generalized medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3099–3107.

You, C.; Zhou, Y.; Zhao, R.; Staib, L.; and Duncan, J. S. 2022. SimCVD: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(9): 2228–2237.

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 Transactions on Medical Imaging*, 39(7): 2531–2540.

Zhao, Y.; Zhong, Z.; Zhao, N.; Sebe, N.; and Lee, G. H. 2022. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *European Conference on Computer Vision*, 535–552.

Zhong, Z.; Zhao, Y.; Lee, G. H.; and Sebe, N. 2022. Adversarial style augmentation for domain generalized urban-scene segmentation. In *Advances in Neural Information Processing Systems*.

Zhou, Z.; Qi, L.; and Shi, Y. 2022. Generalizable medical image segmentation via random amplitude mixup and domain-specific image restoration. In *European Conference on Computer Vision*, 420–436.

Zhou, Z.; Rahman Siddiquee, M. M.; Tajbakhsh, N.; and Liang, J. 2018. UNet++: A nested U-Net architecture for medical image segmentation. In *International Workshop on Deep Learning in Medical Image Analysis*, 3–11.