Domain and Content Adaptive Convolution Based Multi-Source Domain Generalization for Medical Image Segmentation

Shishuai Hu, Zehui Liao, Jianpeng Zhang, and Yong Xia, *Member, IEEE*

***Abstract*— The domain gap caused mainly by variable medical image quality renders a major obstacle on the path between training a segmentation model in the lab and applying the trained model to unseen clinical data. To address this issue, domain generalization methods have been proposed, which however usually use static convolutions and are less flexible. In this paper, we propose a multi-source domain generalization model based on the domain and content adaptive convolution (DCAC) for the segmentation of medical images across different modalities. Specifically, we design the domain adaptive convolution (DAC) module and content adaptive convolution (CAC) module and incorporate both into an encoder-decoder backbone. In the DAC module, a dynamic convolutional head is conditioned on the predicted domain code of the input to make our model adapt to the unseen target domain. In the CAC module, a dynamic convolutional head is conditioned on the global image features to make our model adapt to the test image. We evaluated the DCAC model against the baseline and four state-of-the-art domain generalization methods on the prostate segmentation, COVID-19 lesion segmentation, and optic cup/optic disc segmentation tasks. Our results not only indicate that the proposed DCAC model outperforms all competing methods on each segmentation task but also demonstrate the effectiveness of the DAC and CAC modules. Code is available at https://git.io/DCAC.**

***Index Terms*— Domain generalization, medical image segmentation, dynamic convolution, deep learning.**

I. Introduction

MEDICAL image segmentation is one of the most critical yet challenging steps in computer-aided diagnosis. Since manual segmentation requires considerable expertise and is time-consuming, expensive, and prone to operator-related bias, automated segmentation approaches are in extremely high demand and have been extensively studied [1], [2].

Recent years have witnessed the success of deep learning in medical image segmentation [3], [4], [5]. As a data-driven technique, deep learning requires a myriad amount of annotated training data to alleviate the risk of over-fitting. However, there is usually a small dataset for medical image segmentation tasks, and this relates to the work required in acquiring the images and then, more importantly, in image annotation [6], [7], [8], [9]. Due to the small data issue, the i.i.d. assumption, *i.e.*, each training or test data should be drawn independently from an identical distribution, is less likely to be held. Indeed, the problem of distribution discrepancy between training and test data is particularly severe on medical image segmentation tasks, since the quality of medical images varies greatly over many factors, including different scanners, imaging protocols, and operators [10], [11]. As a result, a segmentation model learned on a set of training images may over-fit the data, and hence has a poor generalization ability on test images, which are collected in another medical center and follow a different distribution. Such undesired performance drop renders a major obstacle on path between the design and clinical application of medical image segmentation tools.

To address this issue, tremendous research endeavors have recently focused on unsupervised domain adaptation (UDA), test time adaptation (TTA), and domain generalization. UDA attempts to alleviate the decrease of generalization ability caused by the distribution shift between the labeled source domain (training) data and unlabelled target domain (test) data in three ways. At the data level, the image-to-image translation is performed to make the quality of source domain data match the quality of target domain data, leading to reduced distribution discrepancy [12], [13], [14]. At the feature level, domain adaptation is achieved by using either adversarial training or feature normalization to extract domain-irrelevant features [15], [16]. At the decision level, various constraints are posed to enforce the consistency between the source domain output and target domain output [17]. Despite their promising performance, UDA methods have a limited clinical value due to the requirement of accessing target domain data [18], [19].

To overcome the limitation of UDA, TTA methods have been proposed to train the segmentation model with the source domain data only, while fine-tuning the trained model with the

Manuscript received 1 August 2022; revised 16 September 2022; accepted 21 September 2022. Date of publication 26 September 2022; date of current version 29 December 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62171377, in part by the National Key Research and Development Program of China under Grant 2022YFC2009903/2022YFC2009900, and in part by the Key Research and Development Program of Shaanxi Province, China, under Grant 2022GY-084. *(Shishuai Hu and Zehui Liao contributed equally to this work.) (Corresponding author: Yong Xia.)*

The authors are with the National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: sshu@mail.nwpu.edu.cn; merrical@mail.nwpu.edu.cn; james.zhang@mail.nwpu.edu.cn; yxia@nwpu.edu.cn).

Digital Object Identifier 10.1109/TMI.2022.3210133

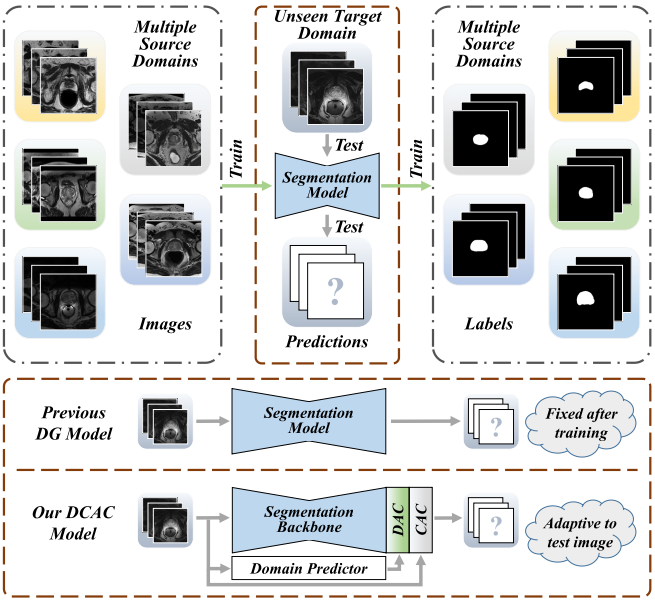

Fig. 1. Illustration of (top) multi-source domain generalization, (middle) previous domain generalization model, and (bottom) proposed DCAC model. The images and corresponding segmentation masks from five source domains are highlighted with different background colors. The **Green** arrows indicate the training process, while the **Gray** arrows highlight the inference process. The segmentation model trained with the data from five source domains is expected to generalize well on the unseen target domain. Previously, a domain generalization model is frozen after training and thus uses the same set of parameters to handle various target domain data. In contrast, our DCAC model can adapt to different test images due to the use of dynamic convolutions. DG: Domain generalization.

target domain data at the test time. It can be accomplished by adding an additional adaptor network to transform [20] or normalize [21] the test data and its features to minimize the domain shift at the test time. Although TTA methods avoid accessing target domain data, they require an extra network to adapt the model to the target data, which increases the spatial and computational complexity.

Domain generalization methods target at boosting the generalization ability of DCNN models and improving their performance in the unseen target domain. An intuitive solution is to extract domain-invariant features via posing domain-invariant constraints to the model or using adversarial training [10], [22], [23]. Nevertheless, it is not easy to differentiate domain-invariant features from domain-specific ones, especially when the target data distribution is completely unknown. To increase the diversity of training data, multiple source domains have been increasingly used to replace the single source domain (see Fig. 1). Multi-source domain generalization methods [11], [24], [25], [26] usually employ meta-learning to minimize the generalization gap between the simulated source domain and target domain. However, if the simulated domain could not cover the unseen target domain, meta-learning-based methods may not perform well. Alternatively, augmentation-based domain generalization methods [27], [28] attempt to simulate the target data distribution via augmenting either the source domain data or the features of source data. Despite their advantages, domain generalization methods still suffer from limited performance, which is attributed mainly to their static nature. Specifically, a domain generalization model is frozen after training and therefore uses the same set of parameters to handle various unseen target data, which have diverse distributions.

In this paper, we propose a multi-source domain generalization model based on the domain and content adaptive convolution (DCAC) for the segmentation of medical images across different modalities. We adopt an encoder-decoder structure as the backbone and design the domain adaptive convolution (DAC) module and content adaptive convolution (CAC) module. To adapt our model to the unseen target domain, the DAC module provides a domain-adaptive head, whose parameters are dynamically generated by the domain-aware controller based on the estimated domain code of the input. To adapt our model to each test image, the CAC module has a content-adaptive head, whose parameters are dynamically produced by the content-aware controller based on the global image features. We have evaluated the proposed DCAC model on three medical image segmentation benchmarks, including the prostate segmentation in MRI scans from six domains, COVID-19 lung lesion segmentation in CT scans from four domains, and optic cup (OC)/optic disc (OD) segmentation in fundus images from four domains.

Our contributions are three-fold.

- We used the domain-discriminative information embedded in the encoder feature maps to generate the domain code of each input image, which establishes the relationship between multiple source domains and the unseen target domain.
- We designed the dynamic convolution-based DAC module and CAC module, which respectively enable our DCAC model to adapt not only to the unseen target domain but also to each test image.
- We presented extensive experimental results, which demonstrate not only the effectiveness of DAC and CAC modules but also the superiority of our DCAC model over state-of-the-art domain generalization techniques on three medical image segmentation tasks.

II. Related Work

A. Domain Generalization for Medical Image Segmentation

Domain generalization methods designed for medical image segmentation can be roughly categorized into augmentation-based, meta-learning-based, and domain-invariant feature learning approaches. **Augmentation-based methods**, such as the deep stacked transformation [27], simulate the distribution of target domain data by augmenting the source domain data. The linear-dependency domain generalization methods [28], [29] perform the augmentation in the feature space, aiming to simulate the distribution of features instead of the distribution of data. With the recent advance of the episodic training strategy for domain generalization in computer vision [30], many **meta-learning-based methods** have been developed to generalize medical image segmentation models to unseen domains [26], [31]. For example, a shape-aware meta-learning scheme [24], which takes the incomplete shape and ambiguous boundary of prediction masks into consideration, was

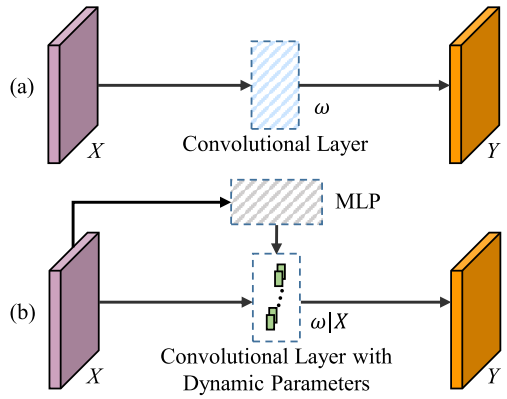

Fig. 2. Comparison between traditional convolution and dynamic convolution: (a) the feature map X is processed by a traditional convolutional layer, whose parameters ω are learned during training; and (b) the feature map X is processed by a dynamic convolutional layer, whose parameters are generated by a multilayer perceptron (MLP) and conditioned on X, *i.e.*, $\omega|X$.

proposed to improve the model generalization for prostate MRI segmentation. In another example, the continuous frequency space interpolation was combined with the episodic training strategy to achieve further performance gains in cross-domain retinal fundus image segmentation and prostate MRI segmentation [11]. Although these methods work well on specific tasks using elaborately tuned parameters, their performance degrades substantially on the target domain when there are only few source domains. Given this, **domain-invariant feature learning methods** [32] have been proposed. Zhao *et al.* [33] adopted domain adversarial learning and mix-up to improve white matter hyperintensity prediction on an unseen target domain. Wang *et al.* [10] built a domain knowledge pool to store domain-specific prior knowledge and then utilized domain attribute to aggregate features from different domains.

Different from these methods, the proposed DCAC model uses dynamic convolutions whose parameters are generated by a controller according to the features of an input image, and thus is able to adapt to the test image from an unknown domain.

B. Dynamic Convolutions

The traditional convolution suffers from limited flexibility, since its parameters ω are learned during training and fixed during inference, regardless of the variations of input, task, and domain (see Fig. 2 (a)). To address this issue, the dynamic convolution has been proposed. Specifically, another network (*e.g.*, an MLP) is employed to generate the convolutional parameters ω based on various conditions (*e.g.*, the input X), and the convolutions with dynamically generated parameters (e.g., $\omega|X$) are then used to process the input (see Fig. 2(b)). Since the parameters ω can change with respect to the current input, task, and/or image domain during inference, the dynamic convolution is far more flexible than its traditional counterpart. Therefore, various dynamic convolutions have been increasingly studied and used in the field of computer vision [34], [35], [36], [37]. A dynamic convolutional layer, in which the filters are generated conditioned on the input image, was proposed for short-range weather prediction based on radar images [38]. The dynamic convolutions, whose parameters are generated conditioned on each target instance, were also integrated to the mask head of an instance segmentation network to improve the accuracy and inference speed [39]. In our previous work, we proposed a convolutional neural network with a dynamic segmentation head, which can be trained on partially labelled abdominal CT scans and be applied to the adaptive segmentation of multiple organs and tumors [40]. In the dynamic head, convolutional parameters are generated conditioned on the combination of a task encode and global image features. By contrast, the DCAC model proposed in this study aims to filter out domain-specific features dynamically and be aware of the content of an input image. Therefore, DCAC contains a DAC head and a CAC head. The DAC head is composed of only one dynamic convolutional layer, in which the dynamic filters are conditioned on the domain code; while the CAC head contains three dynamic convolutional layers, in which the dynamic filters are conditioned on the global features of an input image.

III. Method

A. Problem Definition and Method Overview

Let a set of K source domains be denoted by $D_s = \{(x_{ki}, y_{ki})_{i=1}^{N_k}\}_{k=1}^{K}$, where x_{ki} is the i-th image in the k-th source domain, and y_{ki} is the segmentation mask of x_{ki}. Our goal is to train a segmentation model $F_\theta : x \rightarrow y$ on D_s, which can generalize well to an unseen target domain $D_t = (x_i)_{i=1}^{N_t}$.

The proposed DCAC model is an encoder-decoder structure [3] equipped with a domain predictor, a domain-aware controller, a content-aware controller, and a series of domain-adaptive heads and content-adaptive heads. The workflow of this model consists of four steps. First, the feature map produced by each encoder layer is aggregated using Global Average Pooling (GAP) and concatenated together to be fed to the domain predictor to generate the domain code. Second, based on the generated domain code, the domain-aware controller predicts the parameters of the domain-adaptive head. Third, the content-aware controller uses the final output of the encoder as its input to generate the parameters of the content-adaptive head. Finally, according to the deep supervision strategy, the output of each decoder layer is fed sequentially to a domain-adaptive head and a content-adaptive head, which predict the segmentation result on a pixel-by-pixel basis. The diagram of our DCAC model is shown in Fig. 3. We now delve into its details.

B. Encoder-Decoder Backbone

The backbone used in our DCAC model is a U-shape structure that has an encoder and a decoder, each being composed of $N = 4 \sim 6$ blocks depending on the given segmentation task. Each encoder block contains two convolutional layers with a kernel size of 3, and the first layer has a stride of 2 to downsample the feature map, except for the first encoder block. Each layer is followed by instance normalization and the LeakyReLU activation. In the encoder, the number of

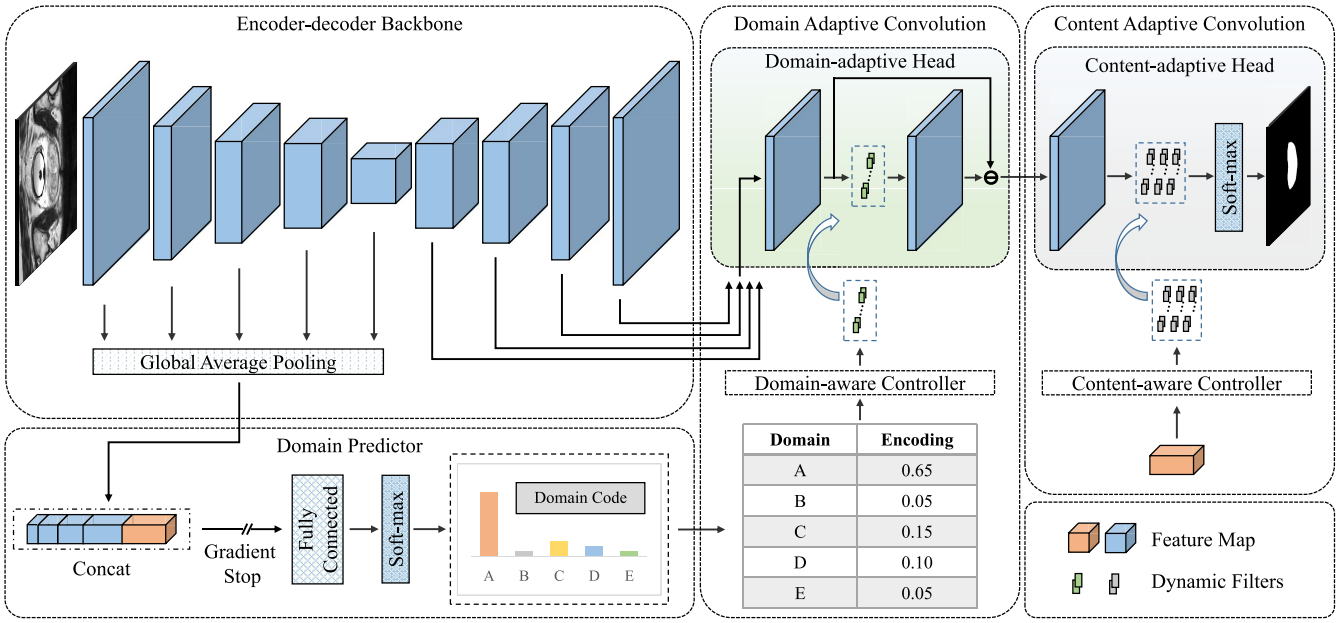

Fig. 3. Architecture of the proposed DCAC model. The feature map in orange color represents $GAP(f_E^N)$, *i.e.*, the output of the N-th encoder block after global average pooling. The traditional convolutions are omitted for simplicity. The dashed boxes with dynamic filters represent dynamic convolutions.

filters is set to 32 in the first layer, then doubled in each next block, and finally fixed with 320 when it becomes larger than 256 [5]. The computation in each encoder block can be formally expressed as

$$f_E^i = Enc^i(f_E^{i-1}; \theta_E^i), \quad i = 1, 2, \cdots, N \tag{1}$$

where θ_E^i represents the parameters of the i-th encoder block Enc^i, f_E^i is the feature map produced by Enc^i, and $f_E^0 = x^i$ is the input image.

Symmetrically, the decoder upsamples the feature map and refines it gradually. In each decoder block, the transposed convolution with a stride of 2 is used to improve the resolution of input feature maps, and the upsampled feature map is concatenated with the corresponding low-level feature map from the encoder before being further processed by two convolutional layers. The computation in each decoder block can be formally expressed as

$$f_D^i = Dec^i(C(f_E^i, U(f_D^{i+1})); \theta_D^i), \ i = N-1, N-2 \cdots, 1 \tag{2}$$

where $U(\cdot)$ represents upsampling, $C(\cdot)$ represents concatenation, θ_D^i represents the parameters of the i-th decoder block Dec^i, f_D^i is the feature map produced by Dec^i, and $f_D^N = f_E^N$.

With this encoder-decoder architecture, multiscale encoder feature maps $\{f_E^i\}_{i=1}^N$ and multiscale decoder feature maps $\{f_D^i\}_{i=1}^{N-1}$ can be generated. It is expected that $\{f_E^i\}_{i=1}^N$ are domain-sensitive and can be utilized to calculate the probabilities of belonging to source domains of the input image. Meanwhile, $\{f_D^i\}_{i=1}^{N-1}$ are expected to be rich-semantic and not subjected to a specific domain, *i.e.*, containing the semantic information of domains and target tasks.

C. Domain Adaptive Convolution

Due to the discrepancy between source domains and the unseen target domain, the encoder-decoder backbone trained with source domain images may not be optimal for target domain images. Therefore, we equipped the backbone with domain-adaptive heads, in which the filters are variable and adaptive to the domain of the input image in the inference stage. For each test image, its probabilities of belonging to source domains, known as a domain code, are calculated by the domain predictor and fed to the domain-aware controller to generate the filters used in the domain-adaptive heads (see Fig. 3).

1) Domain Predictor: Although the target domain is not identical to each source domain, an image in the target domain may similar to those in one or more source domains. And such 'domain attribute' of the image can be used as the clue to guide the adaptive processing of it. Therefore, we design the domain predictor to predict the probability of each target domain image belonging to each source domain.

The domain predictor takes multi-scale encoder feature maps $\{f_E^i\}_{i=1}^N$ as its input. Each feature map f_E^i is aggregated with GAP, and the aggregated features at all scales are then concatenated into a vector. To predict the domain code of the input image, the vector is fed to a classification module, which is composed of a fully-connected layer $FC(\cdot)$ and a soft-max layer $SM(\cdot)$. The calculation of each domain code can be formally expressed as

$$\mathscr{D}^p = SM((FC(C(GAP(f_E^1), \cdots, GAP(f_E^N))); \theta_{FC})), \tag{3}$$

where θ_{FC} represents the parameters of $FC(\cdot)$. The domain code $\mathscr{D}^p$ is a K-dimensional vector that satisfies $\sum_{k=1}^K \mathscr{D}_k^p = 1$. During training, since each input image is sampled from one of K source domains, the ground truth

domain code that supervises the training of domain predictor is a one-hot K-dimensional vector. Note that image segmentation and domain prediction are different tasks, though using the same set of features extracted by encoder blocks. To avoid the interference with the image segmentation performance caused by domain prediction, we adopt the gradients truncation strategy to stop the gradients back propagated from the fully-connected layer in the domain predictor (see Fig. 3).

2) Domain-Aware Controller: We use a single traditional convolutional layer as the domain-aware controller $\phi_d(\cdot)$, which maps the domain code to the parameters ω_d of the filters in the domain adaptive head. Such mapping can be formally expressed as

$$\omega_d = \phi_d(\mathscr{D}^p; \theta_\phi^d) \tag{4}$$

where θ_ϕ^d represents the parameters in this controller.

3) Domain-Adaptive Head: A lightweight domain-adaptive head is designed to enable dynamic convolutions, which are responsive to specific domains. This head contains a traditional convolutional layer and a dynamic convolutional layer, both using filters with a kernel size of 1. The traditional layer reduces the channels of the input feature map to $K \times C$, where C is the number of segmentation classes. Since there exists a skip connection to enforce residual learning, the output of the dynamic layer has $K \times C$ channels, too. Therefore, there are totally $(K \times C)^2 + (K \times C)$ parameters in the dynamic layer, which are generated dynamically by the domain-aware controller $\phi_d(\cdot)$ conditioned on the domain code $\mathscr{D}^p$. Thanks to the superiority of dynamic convolutions, the parameters in the dynamic layer are domain-specific, and the output of the dynamic layer can represent the domain-specific feature.

To accelerate the convergence of our DCAC model, we adopt the multi-scale supervision strategy. Given the feature map f_D^i generated by the i-th decoder block, the output of the domain-adaptive head is computed as

$$f_{DAC}^i = Conv_O^i(f_D^i) - Conv_O^i(f_D^i) * \omega_d, \quad i = N-1, N-2 \cdots, 1 \tag{5}$$

where $*$ represents the convolution, and $Conv_O^i(\cdot)$ is the traditional convolutional layer.

D. Content Adaptive Convolution

The proposed DCAC model is expected to adapt not only to the unseen test domain but also to each test image. Therefore, we equipped our segmentation backbone with content adaptive convolutions, which are implemented using a content-adaptive head whose parameters are generated dynamically by a content-aware controller.

1) Content-Aware Controller: The content-aware controller is a traditional convolutional layer, denoted by ϕ_c. The input of this controller is the global image representation, which is the feature map generated by the encoder (*i.e.*, the output f_E^N of the N-th encoder block) and aggregated by global average pooling. The output is the ensemble of parameters of the content-adaptive head, which can be formally expressed as

$$\omega_c = \phi_c(GAP(f_E^N); \theta_\phi^c) \tag{6}$$

where θ_ϕ^c represents the parameters of the controller ϕ_c.

2) Content-Adaptive Head: The content-adaptive head, which is placed after the domain-adaptive head, contains three stacked dynamic convolutional layers using filters with a kernel size of 1. The first two layers have $K \times C$ channels, and the last layer has C channels. Thus there are totally $2 \times ((K \times C)^2 + (K \times C)) + ((K \times C) \times C + C)$ dynamic parameters in this head. These parameters, denoted by $\omega_c = \{\omega_{c1}, \omega_{c2}, \omega_{c3}\}$, are generated by the controller ϕ_c according to the globally aggregated image feature map f_E^N.

The content-adaptive head uses the output of domain-adaptive head f_{DAC}^i as its input. This head acts as a pixel classifier, performing image segmentation via predicting class labels on a pixel-by-pixel basis. The computation of segmentation result p^i can be formally expressed as

$$p^i = SM(((f_{DAC}^i * \omega_{c1}) * \omega_{c2}) * \omega_{c3}), \quad i = N-1, N-2, \cdots, 1 \tag{7}$$

where $SM(\cdot)$ represents the soft-max operation.

E. Training and Test

1) Training: Besides image segmentation, the proposed DCAC model also performs domain classification using the domain predictor. For the classification task, the objective is the cross-entropy loss, which can be calculated as

$$\mathcal{L}_{cls} = -\sum_{k=1}^{K} d_k \log(d_k^p) \tag{8}$$

where d_k is the domain label, and d_k^p is the soft-max probability of belonging to the k-th domain.

For the segmentation task, the Dice loss and cross-entropy loss are used jointly as the objective. The segmentation loss at each scale can be calculated as

$$\mathcal{L}_{seg}^i = 1 - \frac{2\sum_{v=1}^{V} p_v^i y_v^i}{\sum_{v=1}^{V}\left(p_v^i + y_v^i + \epsilon\right)} - \sum_{v=1}^{V}\left(y_v^i \log p_v^i + \left(1 - y_v^i\right)\log\left(1 - p_v^i\right)\right) \tag{9}$$

where p_v^i and y_v^i denote the prediction and ground truth of the v-th voxel in the output of the i-th decoder block, V represents the number of voxels, and ϵ is a smooth factor to avoid dividing by 0.

Since deep supervision is used, the total loss is defined as follows

$$\mathcal{L} = \mathcal{L}_{cls} + \sum_{i=1}^{N-1} \omega^i \mathcal{L}_{seg}^i \tag{10}$$

where ω^i is a weighting vector that enables higher resolution output to contribute more to the total loss [5].

TABLE I
STATISTICS OF THREE DATASETS USED FOR THIS STUDY

Task	Modality	Number of Domains	Cases in Each Domain	Total Cases
Prostate Segmentation	MRI	6	30; 30; 19; 13; 12; 12	116
COVID-19 Segmentation	CT	4	28; 19; 58; 15	120
OC/OD Segmentation	Color Fundus Image	4	50/51; 99/60; 320/80; 320/80*	789/281*

* Data split (training/test cases) was provided by [10].

2) Test: During inference, given a test image x, the multiscale encoder feature maps $\{f_E^i\}_{i=1}^N$ and multiscale decoder feature maps $\{f_D^i\}_{i=1}^{N-1}$ can be produced by the trained encoder-decoder backbone. Based on $\{f_E^i\}_{i=1}^N$, the trained domain predictor can generate a K-dimensional domain code. Based on the code, the trained domain-aware controller can generate the parameters for the domain-adaptive head. Meanwhile, based on the feature map produced by the last encoder block (*i.e.*, f_E^N), the content-aware controller can generate the parameters for the content-adaptive head. Finally, the feature map produced by the decoder is fed sequentially to the domain-adaptive dynamic head and content-adaptive head to generate the segmentation result. Note that deep supervision is carried out only in the training stage and the segmentation is not performed at coarse scales in the test stage.

IV. EXPERIMENTS

We evaluated the proposed DCAC model against the baseline and state-of-the-art domain generalization models on three tasks, including prostate segmentation using MRI, COVID-19 lesion segmentation using CT, and OC/OD segmentation using color fundus image. These tasks cover different image modalities and represent variable domain shifts in cross-domain medical image segmentation problems.

A. Datasets

Three datasets were used for this study. For prostate segmentation, the dataset contains 116 T2-weighted MRI cases from six domains [24], [41], [42], [43]. Following [24] and [11], we preprocessed the MRI data and only preserved the slices with the prostate region for consistent and objective segmentation evaluation. For COVID-19 lesion segmentation, the dataset consists of 120 RT-PCR positive CT scans with pixel-level lesion annotations, collected from the first multi-institutional, multi-national expert annotated COVID-19 image database [18], [44], [45]. For OC/OD segmentation, the dataset contains 789 cases for training and 281 cases for test, which are collected from four public fundus image datasets and have inconsistent statistical characteristics [10], [46], [47], [48]. The statistics of three datasets were summarized in Table I.

B. Implementation Details

The images in each segmentation task were normalized by subtracting the mean and dividing by the standard deviation. To make a compromise between the network complexity and input image size, the mini-batch size was set to 32 for 2D prostate segmentation with a patch size of 256×256, set to 16 for 2D OC/OD segmentation with a patch size of 512×512, and set to 2 for 3D COVID-19 lesion segmentation with a patch size of $128 \times 196 \times 196$. To expand the training set, several data augmentation techniques were used, including random cropping, rotation, scaling, flipping, adding Gaussian noise, and elastic deformation. The SGD algorithm with a momentum of 0.99 was adopted as the optimizer. The initial learning rate lr_0 was set to 0.01 and decayed according to the polynomial rule $lr = lr_0 \times (1 - t/T)^{0.9}$, where t is the current epoch and T is the maximum epoch. The maximum epoch T was set to 200 for 2D prostate segmentation, 500 for 2D OC/OD segmentation, and 1000 for 3D COVID-19 lesion segmentation. Our DCAC was implemented using the PyTorch framework on a workstation with a NVIDIA 2080Ti GPU.

C. Comparative Experiments and Analysis

We compared the proposed DCAC model with the 'Intra-domain' setting, 'DeepAll' baseline, and four domain generalization methods including (1) a data-augmentation based method called BigAug [27], (2) two meta-learning methods called SAML [24] and FedDG [11], and (3) a domain-invariant feature learning approach called DoFE [10]. Under the 'Intra-domain' setting, training and test data are from the same domain and the three-fold cross-validation is used. Whereas under the 'DeepAll' setting, the model is trained on the data aggregated from all source domains and tested directly on the unseen target domain. Note that the backbone, *i.e.*, nnUNet, was kept as the same for all these methods in all experiments unless otherwise indicated. For each segmentation task, the leave-one-domain-out strategy was used to evaluate the performance of each domain generalization method, *i.e.*, training on $K - 1$ source domains and evaluating on the left unseen target domain. Each domain is chosen as the target domain in turn. The segmentation performance was measured by the Dice Similarity Coefficient (DSC) and Average Surface Distance (ASD). The DSC (%) and ASD (pixel) characterize the accuracy of predicted masks and boundaries, respectively.

1) Comparative Results in Prostate Segmentation: Table II gives the DSC and ASD values obtained by our DCAC model and six segmentation models in each target domain and the average performance over six domains. As expected, the performance of DeepAll seems to be worse on average than that of Intra-domain, due to the distribution discrepancy between the source (training) data and target (test) data. Meanwhile, it shows that the augmentation-based method BigAug performs worse than meta-learning-based methods (*i.e.*, SAML and FedDG), indicating that simply augmenting training data is insufficient to simulate the data distribution

TABLE II
PERFORMANCE OF OUR DCAC MODEL AND SIX SEGMENTATION MODELS IN PROSTATE SEGMENTATION. THE BEST RESULTS EXCEPT FOR THE RESULTS OF INTRA-DOMAIN ARE HIGHLIGHTED WITH **BOLD**

Models	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 [27]	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 [24]	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 [11]	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 [10]	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
Ours (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**

* Standard deviation and significance test can be found at https://arxiv.org/abs/2109.05676.

TABLE III
PERFORMANCE OF OUR DCAC MODEL AND SIX SEGMENTATION MODELS IN COVID-19 LESION SEGMENTATION. THE BEST RESULTS EXCEPT FOR THE RESULTS OF INTRA-DOMAIN ARE HIGHLIGHTED WITH **BOLD**

Models	Domain 1		Domain 2		Domain 3		Domain 4		Average	
	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓
Intra-domain	62.49	21.05	51.34	21.08	70.49	6.02	62.01	8.14	61.58	14.07
DeepAll	63.09	19.13	60.87	19.44	66.40	12.21	62.57	9.39	63.23	15.04
BigAug [27]	63.55	18.09	59.57	19.53	67.19	13.20	64.39	9.39	63.68	15.05
SAML [24]	63.98	15.96	61.39	18.97	67.19	12.87	65.38	9.39	64.49	14.30
FedDG [11]	63.97	17.68	60.88	17.85	66.96	13.10	64.98	9.30	64.20	14.48
DoFE [10]	**64.76**	**12.43**	61.11	18.56	67.46	11.74	65.05	9.71	64.60	13.11
Ours (DCAC)	64.03	17.05	**62.52**	**15.38**	**67.87**	**10.63**	**65.96**	**7.98**	**65.10**	**12.76**

* Standard deviation and significance test can be found at https://arxiv.org/abs/2109.05676.

TABLE IV
PERFORMANCE (OC, OD) OF OUR DCAC MODEL AND SIX SEGMENTATION MODELS IN OC/OD SEGMENTATION. THE BEST RESULTS EXCEPT FOR THE RESULTS OF INTRA-DOMAIN ARE HIGHLIGHTED WITH **BOLD**

Models	Domain 1		Domain 2		Domain 3		Domain 4		Average	
	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓	DSC↑	ASD↓
Intra-domain	(80.06, 95.82)	(20.13, 7.53)	(73.13, 87.79)	(24.91, 18.75)	(83.80, 93.20)	(11.20, 9.64)	(84.46, 93.41)	(8.99, 7.51)	86.46	13.58
DeepAll	(79.04, 95.82)	(20.32, 7.63)	(73.02, 87.34)	(24.99, 18.70)	(82.26, 91.37)	(12.01, 11.40)	(84.85, 92.27)	(8.39, 7.83)	85.75	13.91
BigAug [27]	(80.37, 95.59)	(19.50, 7.75)	(74.73, 87.40)	(22.64, 18.89)	(85.39, 92.04)	(10.07, 11.09)	(86.47, 93.05)	(8.32, 7.75)	86.88	13.25
SAML [24]	(81.03, 95.74)	(19.31, 7.66)	(76.61, 87.29)	(19.31, 19.20)	(85.40, 93.92)	(9.99, 8.62)	(86.06, 94.76)	(8.86, 5.90)	87.60	12.36
FedDG [11]	(81.66, 95.47)	(18.79, 7.81)	(76.31, 86.34)	(19.98, 19.57)	(85.23, 93.36)	(10.86, 9.12)	(85.27, 94.68)	(8.94, 6.02)	87.29	12.64
DoFE [10]	(**81.95**, 96.04)	(**18.59**, 7.05)	(**78.31**, **89.20**)	(**16.40**, **15.75**)	(85.51, 93.23)	(10.06, 9.76)	(86.61, 94.28)	(8.28, 6.99)	88.14	11.61
Ours (DCAC)	(81.43, **96.54**)	(19.20, **6.35**)	(77.72, 87.85)	(17.15, 18.28)	(**86.80**, **94.28**)	(**9.14**, **8.11**)	(**87.68**, **95.40**)	(**7.12**, **5.20**)	**88.47**	**11.32**

* Standard deviation and significance test can be found at https://arxiv.org/abs/2109.05676.

of the target domain. It also shows that DoFE is superior to FedDG but slightly inferior to SAML, suggesting that the domain-invariant feature learning approach (*i.e.*, DoFE) can disentangle domain-sensitive features, but it can hardly adapt to different domain discrepancies automatically. More importantly, it reveals that the proposed DCAC mode not only beats Intra-domain and DeepAll but also outperforms four state-of-the-art domain generalization methods. We believe the superior performance can be attributed to the fact that, with dynamic convolution, our model is capable of adapting to both the predicted domain code and extracted global features of the input image.

2) Comparative Results in COVID-19 Lesion Segmentation and OC/OD Segmentation: The segmentation performance of our DCAC model and six segmentation models on the COVID-19 lesion segmentation task and OC/OD segmentation task was given in Table III and Table IV, respectively. In COVID-19 lesion segmentation, the average DSC of Intra-domain is surprisingly worse than that of DeepAll. A possible reason is that the amount of training data in a single domain (see Table I) is far from sufficient for training a DCNN model, leading to serious over-fitting of the small training dataset. By contrast, aggregating the data in multiple domains can benefit model training and thus results in improved performance. Meanwhile, it seems that the domain-invariant feature learning method is relatively better than meta-learning methods in both experiments, indicating the sensitivity of meta-learning-based methods to the number of source domains. When there are less source domains, the diversity of the generalization gap simulated by meta-learning is highly restricted. Comparing to these methods, our model is less susceptible to the number of source domains and achieves

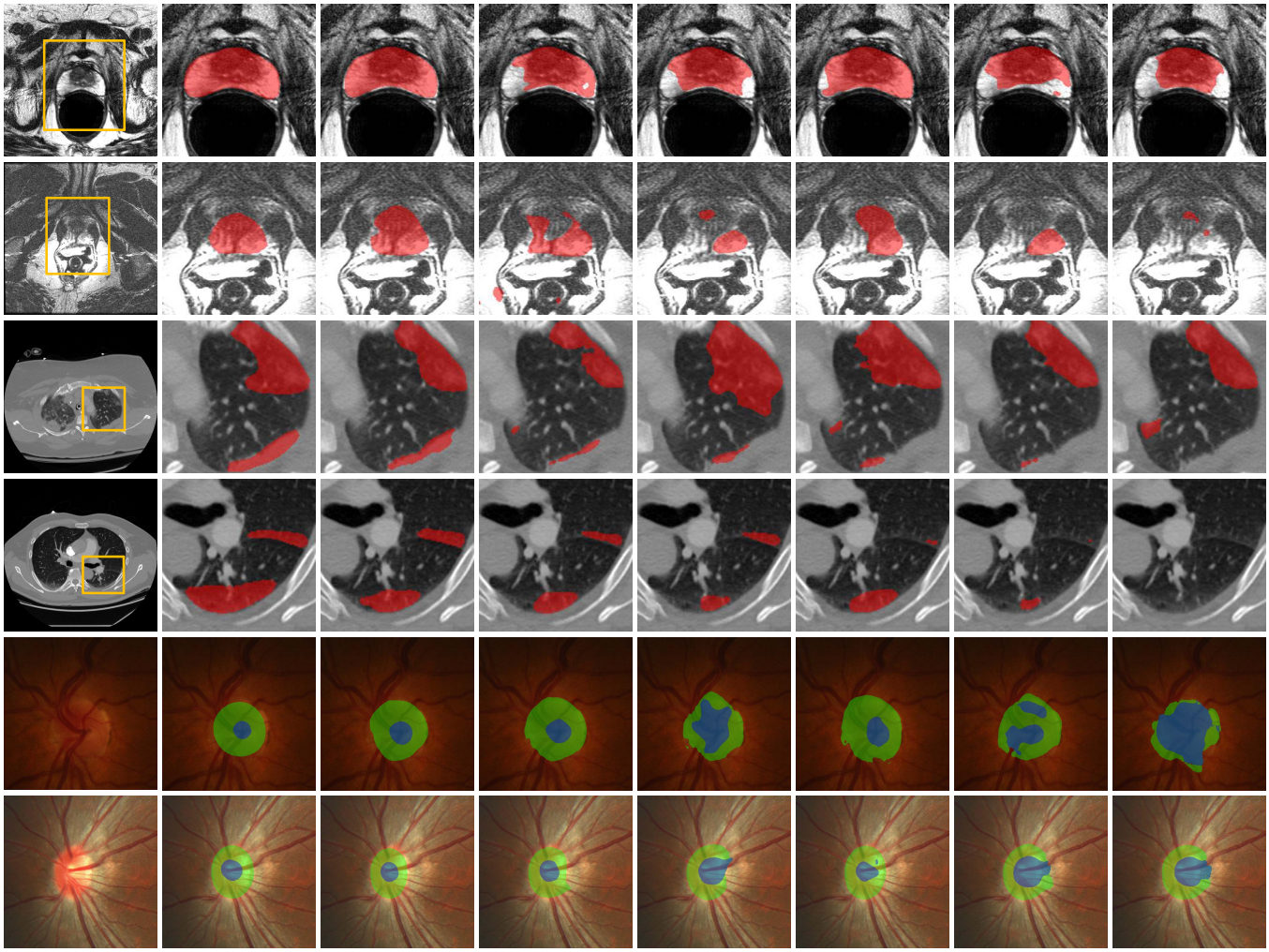

Fig. 4. Visualization of the results predicted by ours (DCAC) and five competing methods on the three segmentation tasks, together with ground truth. Best viewed in color.

stable performance gain on both segmentation tasks. This observation is consistent with what we observed in Table II.

3) Visualization Results of DCAC and Other Competing Methods: We visualized the segmentation results of our DCAC and four competing domain generalization methods in Fig. 4. We also displayed the results of DeepAll and ground truth for reference. It shows that the segmentation results produced by our DCAC model are the most similar to the ground truth over all three segmentation tasks, which confirms the effectiveness of our DCAC model against the state-of-the-art in the three generalizable medical image segmentation benchmarks with different imaging modalities.

V. Discussion

The prostate segmentation task was chosen as a case study, and ablation studies were conducted on this task to investigate the effectiveness of newly designed DAC and CAC modules and the domain-discriminatory ability of extracted features.

A. Ablation Analysis

In this work, we designed the DAC module and CAC module to make our model capable of adapting to the unseen test domain and test image, respectively. To evaluate the contributions of these two modules, we compared our model with its variant that uses only one module. Meanwhile, we changed the order between DAC and CAC for comparison, denoted as CDAC. We also compared a variant, denoted by D-CAC, that uses the concatenation of domain code and global image features to generate one and only one unified dynamic head. The performance of DeepAll (Baseline), our DCAC model, and its variants was given in Table V. It shows that CDAC achieves similar performance to our DCAC, and both of them outperform not only D-CAC but also the variant without either DAC or CAC. The results confirm that either DAC or CAC contributes to the final results and the two-dynamic-head strategy is superior to the unified dynamic head.

We visualized the segmentation results of DCAC and three variants in Fig. 5. We also displayed the results of DeepAll and ground truth for reference. It shows that our DCAC model can produce more accurate segmentation results of unseen test images, particularly in the boundary region.

B. Domain-Discriminatory Power of Extracted Features

In our DCAC model, the DAC module relies heavily on the domain code $\mathscr{D}^p$ predicted based on image features. To predict

TABLE V
PERFORMANCE OF DEEPALL, OUR DCAC MODEL, AND ITS SIX VARIANTS IN PROSTATE SEGMENTATION

Models	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↓
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
D-CAC	91.24	1.37	89.94	0.92	86.72	1.67	89.23	1.34	79.51	3.54	89.90	0.96	87.74	1.70
Ours w/o DAC	91.13	1.12	89.62	1.01	84.75	2.17	89.31	1.48	80.79	2.11	89.93	0.93	87.59	1.47
Ours w/o CAC	91.69	1.01	89.96	0.97	85.27	1.89	89.19	1.33	78.44	2.35	90.65	0.90	87.53	1.41
CDAC	91.74	1.11	90.72	0.90	86.05	2.15	89.18	1.52	83.27	2.02	90.21	0.94	88.53	1.44
$DC^{(p)}AC$	90.02	1.61	89.78	0.96	85.44	1.94	88.28	2.04	81.37	2.34	90.41	0.88	87.55	1.63
DCAC-NG	90.67	1.29	88.12	1.04	78.58	2.87	88.31	1.41	81.18	1.97	89.69	1.01	86.09	1.60
Ours (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

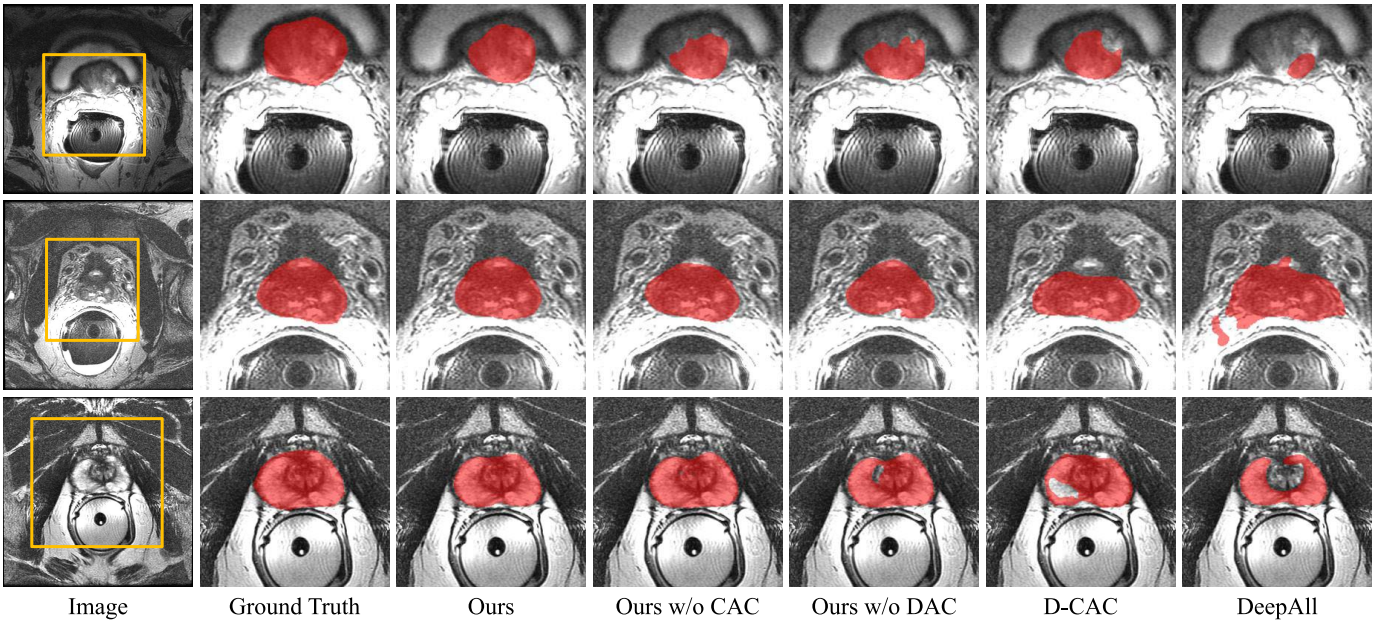

Fig. 5. Visualization of one slice from each of three prostate MRI scans, corresponding segmentation ground truth, and the results obtained by applying our DCAC, three variants of DCAC, and DeepAll.

$\mathscr{D}^p$ accurately, the features should have sufficient domain-discriminatory power. Instead of using the single-scale global feature produced by the last encoder block, we chose the multi-scale global features extracted by the encoder at all scales for our study (see Fig. 3). To verify the superiority of our multi-scale features, we compared the domain classification accuracy achieved by using each of these two types of features. The obtained confusion matrices were visualized in Fig. 6. It shows that using multi-scale features can produce more accurate domain classification than using single-scale features, suggesting that the multi-scale feature maps produced by encoder blocks contain domain-specific information and can be used to generate the domain code.

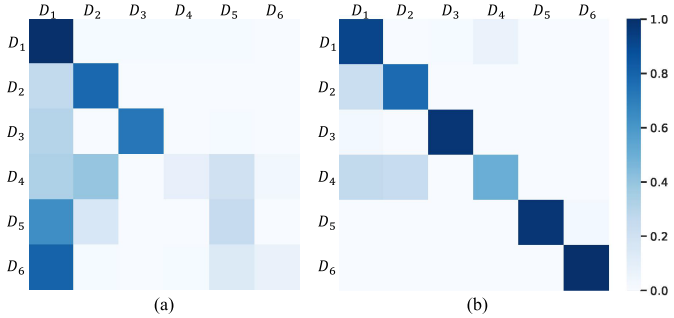

Fig. 6. Confusion matrices of domain classification achieved by using (a) single-scale global features or (b) multi-scale global features as input of domain predictor. In each confusion matrix, the rows and columns represent true and predicted domains, respectively, and each element (i, j) represents the probability of predicting domain D_i as D_j. Particularly, a diagonal element represents the true positive rate of predicting the corresponding domain.

C. Analysis of Complexity

Besides the segmentation backbone, the proposed DCAC model also contains a domain predictor, two controllers, and two dynamic convolutional heads, which, fortunately, all have lightweight structures. Therefore, apart from the parameters in the backbone, DCAC just has a few extra parameters and consumes a little extra training time. We chose the prostate segmentation task as a case study and listed the number of parameters, GFLOPs, model size, and training time cost of our DCAC model and two state-of-the-art domain generalization methods (*i.e.*, DoFE and SAML) in Table VI. Note that the

TABLE VI
NUMBER OF PARAMETERS, GFLOPS, MODEL SIZE, AND TRAINING DURATION TIME COST OF DIFFERENT MODELS IN PROSTATE SEGMENTATION

Models	#Parameters ($\times 10^6$)	GFLOPs	Model Size (MB)	Training Time Cost (Hours)
DoFE [10]	30.1	32.7	145.4	7.1
SAML [24]	30.0	32.4	119.9	11.8
Ours (DCAC)	30.1	32.5	120.5	6.2

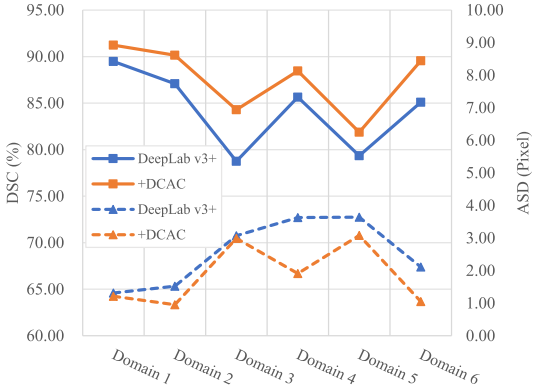

Fig. 7. Performance of DeepLab V3+ with and without DCAC on prostate segmentation task. The values in square represent DSC, and the values in triangle denote ASD.

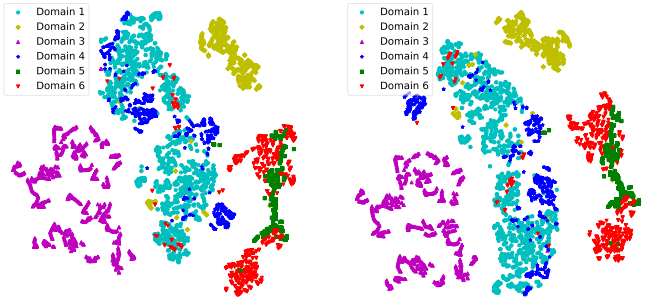

Fig. 8. Visualization of feature maps of input image (right) with and (left) without perturbations in 2D using t-SNE. The points in different colors represent the images from different domains.

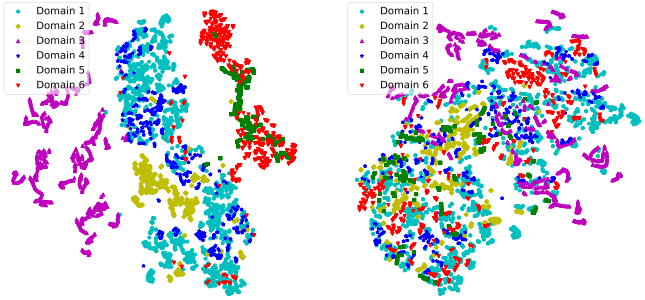

Fig. 9. t-SNE visualization of features before (left) and after (right) being processed by the DAC head. The points in different colors represent images from different domains.

parameters of the entire model (including the backbone) were counted for all methods, and the backbone was also taken into consideration when calculating GFLOPs and the model size. It shows that, although three models have a similar number of parameters and GFLOPs, the size of DCAC model is significantly smaller than DoFE, since DoFE uses an additional domain knowledge pool to store domain prior knowledge for domain-sensitive feature matching during inference, but our DCAC achieves this using the light-weighted dynamic convolutions. Moreover, our DCAC has much less training time cost than DoFE and SAML. The extremely high time cost of SAML can be attributed to the fact that SAML relies heavily on meta-train and meta-test to update model parameters in each iteration, which is time-consuming. As for DoFE, it requires to update the domain knowledge pool and perform feature embedding during each training step. In summary, our results indicate that, comparing to DoFE and SAML, the proposed DCAC model is able to produce more accurate segmentation results with less spatial and computational complexity.

D. Applying to Other Backbone

The DCAC is proposed following a modular design and can be incorporated into other encoder-decoder backbones to improve their performance. To validate this, we adopted DeepLab V3+ [49] as the segmentation backbone and incorporated DCAC into it by adding a domain predictor and two dynamic heads. We carried out the prostate segmentation task again and compared the performance of DeepLab V3+ with and without DCAC on six domains in Fig. 7. It reveals that plugging DCAC into DeepLab V3+ results in performance gains on all domains, evidenced by the consistent increase of DSC and decrease of ASD. On average, the mean DSC improves from 84.23% to 87.59%, and the mean ASD drops from 2.55 to 1.87. The results are consistent with those reported in Table II, confirming the usability and effectiveness of our DCAC again.

E. Generalization Analysis of DAC and CAC

To make the proposed DCAC model generalizable to unseen target domains, the domain adaptive head itself should be generalizable. To validate this, we chose the model trained using Domain 1 $\sim$ Domain 5 on the prostate segmentation task as a case study. The feature map of each input image (including both training and test cases) produced by the DAC module, denoted by $Conv_O^1(f_D^1) * \omega_d$, is visualized in 2D using t-SNE (see the left part of Fig. 8). In the meantime, we added the white Gaussian noises with a standard deviation of 0.2 as perturbations to each input image and visualized the feature map produced by DAC in the right part of Fig. 8. It shows that adding perturbations to input images leads to little impact on the topology of the DAC features. Therefore, the feature extraction performed by the DAC module is robust to input perturbations, suggesting that the DAC module does not over-fit the data from source domains.

To demonstrate that the DAC convolutions ω_d can filter domain-specific features, we also visualized the image features before and after being filtered by ω_d in 2D using t-SNE (see Fig. 9). It shows that, after being filtered by ω_d, the image features from different domains, which previously can be largely separated from each other, become indistinguishable.

It indicates that the DAC module can effectively filter out domain-specific features.

The dynamic convolutions in the CAC module, whose parameters ω_c are dynamically produced based on the global features of each input image, aim to adapt our DCAC model to the input. To validate the effectiveness of these convolutions, we shuffled the ω_c generated from different input images and evaluated the impact of such perturbations on the segmentation performance. The segmentation performance of our DCAC model with (denoted by DC$^{(p)}$AC) or without ω_c perturbations on the prostate dataset was given in Table V. It shows that mutating ω_c deteriorates the performance of our DCAC model on each unseen target domain, decreasing the average DSC from 88.61% to 87.55%. It confirms that the ω_c estimated by the content-aware controller is suitable for each input image.

F. Gradient Truncation in Domain Predictor

Besides image segmentation, the proposed DCAC model also performs domain classification using the domain predictor. These two tasks share the features extracted from the same encoder. For the segmentation task, the extracted features should be domain-insensitive so that the segmentation performance would be less affected by the domain discrepancy. Whereas the classification task requires the extracted features to have domain-discriminatory power. To make a compromise between the domain classification accuracy and the segmentation performance, we adopt the gradients truncation strategy in the domain predictor. Thus, only the parameters in the fully-connect layer in the domain predictor can be optimized for domain classification. It can be observed from Fig. 6 that the domain attributions can still be largely discriminated when adopting the gradient truncation strategy. We also analyzed the segmentation performance when using (DCAC) and not using (DCAC-NG) gradients truncation. The quantitative results were shown in Table V. It reveals that the overall segmentation performance on the unseen target domains can be dramatically decreased when the shared encoder is optimized for both domain classification and segmentation. It confirms the benefit brought by the gradient truncation strategy.

VI. Conclusion

This paper proposes a multi-source domain generalization model called DCAC, which uses two dynamic convolutional heads. One dynamic head is conditioned on the predicted domain code of the input to make the DCAC model adapt to the target domain, while the other dynamic head is conditioned on global image features to make the model adapt to the input image. Our results on the prostate segmentation, COVID-19 lesion segmentation, and OC/OD segmentation tasks suggest that, after training on the data from multiple source domains, the proposed DCAC model can generalize well on an unseen target domain, achieving improved average performance over the baseline and four state-of-the-art domain generalization methods.

However, the proposed DCAC model still has two limitations. First, it is designed for multi-source domain generalization, and therefore cannot be directly applied to single-source domain generalization [50], [51]. In other words, it requires training data from multiple source domains. Nevertheless, we believe multi-source domain generalization is a promising research direction orthogonal to existing single-source domain generalization methods. Second, the performance gain achieved by our DCAC on CT images is less than that on MR and fundus images (see Table II, Table III, and Table IV). It can be attributed to the fact that CT images for the same phase generally have more consistent image quality, whereas MR and fundus images hold more vendor-specific variations [27]. In our future work, we will extend the proposed DCAC model to multi-source, multi-modality, and multi-task scenarios, aiming to provide a large-scale pre-trained segmentation model for various downstream medical image segmentation tasks.

Acknowledgment

The authors acknowledge the RSNA and Society of Thoracic Radiology (STR), the European Society of Medical Imaging Informatics, the American College of Radiology, and the American Association of Physicists in Medicine, and their critical role in the creation of the free publicly available RICORD dataset used for this study. They also appreciate the efforts devoted by the authors of [10] and [24] to collect and share the prostate MR and fundus imaging data for comparing generalizable medical image segmentation algorithms.

References

[1] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.
[2] X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, and S. Yu, "A survey on incorporating domain knowledge into deep learning for medical image analysis," *Medical Image Anal.*, vol. 69, Apr. 2021, Art. no. 101985.
[3] T. Falk *et al.*, "U-Net: Deep learning for cell counting, detection, and morphometry," *Nature Methods*, vol. 16, no. 1, pp. 67–70, Jan. 2019.
[4] Z. Zhou *et al.*, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Dec. 2020.
[5] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Dec. 2020.
[6] R. Guo, M. Pagnucco, and Y. Song, "Learning with noise: Mask-guided attention model for weakly supervised nuclei segmentation," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Cham, Switzerland: Springer, 2021, pp. 461–470.
[7] X. Wang, H. Chen, H. Xiang, H. Lin, X. Lin, and P.-A. Heng, "Deep virtual adversarial self-training with consistency regularization for semi-supervised medical image classification," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 102010.
[8] Y. Huang *et al.*, "Noise-powered disentangled representation for unsupervised speckle reduction of optical coherence tomography images," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2600–2614, Oct. 2021.
[9] Y. Yang *et al.*, "Towards unbiased COVID-19 lesion localisation and segmentation via weakly supervised learning," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2021, pp. 1966–1970.
[10] S. Wang, L. Yu, K. Li, X. Yang, C.-W. Fu, and P.-A. Heng, "DoFE: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 4237–4248, Dec. 2020.
[11] Q. Liu, C. Chen, J. Qin, Q. Dou, and P.-A. Heng, "FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1013–1023.
[12] Y. Yang and S. Soatto, "FDA: Fourier domain adaptation for semantic segmentation," in *Proc. CVPR*, Jun. 2020, pp. 4085–4095.

[13] C. Chen, Q. Dou, H. Chen, J. Qin, and P. A. Heng, "Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2494–2505, Jul. 2020.

[14] D. Liu *et al.*, "PDAM: A panoptic-level feature alignment framework for unsupervised domain adaptive instance segmentation in microscopy images," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 154–165, Jan. 2021.

[15] Q. Liu, Q. Dou, L. Yu, and P. A. Heng, "MS-Net: Multi-site network for improving prostate segmentation with heterogeneous MRI data," *IEEE Trans. Med. Imag.*, vol. 39, no. 9, pp. 2713–2724, Sep. 2020.

[16] Y. Shen *et al.*, "Domain-invariant interpretable fundus image quality assessment," *Med. Image Anal.*, vol. 61, Apr. 2020, Art. no. 101654.

[17] S. Wang, L. Yu, K. Li, X. Yang, C.-W. Fu, and P.-A. Heng, "Boundary and entropy-driven adversarial learning for fundus image segmentation," *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Cham, Switzerland: Springer, 2019, pp. 102–110.

[18] E. B. Tsai *et al.*, "The RSNA international COVID-19 open radiology database (RICORD)," *Radiology*, vol. 299, no. 1, pp. 204–213, Apr. 2021.

[19] H. Roth *et al.*, "Rapid artificial intelligence solutions in a pandemic—The COVID-19–20 lung CT lesion segmentation challenge," *Med. Image Anal.*, vol. 82, Nov. 2022, Art. no. 102605.

[20] Y. He, A. Carass, L. Zuo, B. E. Dewey, and J. L. Prince, "Autoencoder based self-supervised test-time adaptation for medical image analysis," *Med. Image Anal.*, vol. 72, Aug. 2021, Art. no. 102136.

[21] N. Karani, E. Erdil, K. Chaitanya, and E. Konukoglu, "Test-time adaptable neural networks for robust medical image segmentation," *Med. Image Anal.*, vol. 68, Feb. 2021, Art. no. 101907.

[22] X. Fan, Q. Wang, J. Ke, F. Yang, B. Gong, and M. Zhou, "Adversarially adaptive normalization for single domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8208–8217.

[23] Q. Zhou, W. Zhou, S. Wang, and Y. Xing, "Duplex adversarial networks for multiple-source domain adaptation," *Knowl.-Based Syst.*, vol. 211, Jan. 2021, Art. no. 106569.

[24] Q. Liu, Q. Dou, and P.-A. Heng, "Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Cham, Switzerland: Springer, 2020, pp. 475–485.

[25] Y. Du, X. Zhen, L. Shao, and C. G. M. Snoek, "MetaNorm: Learning to normalize few-shot batches across domains," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–23.

[26] X. Liu, S. Thermos, A. O'Neil, and S. Tsaftaris, "Semi-supervised meta-learning with disentanglement for domain-generalised medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Cham, Switzerland: Springer, Jun. 2021.

[27] L. Zhang *et al.*, "Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2531–2540, Jul. 2020.

[28] H. Li, Y. Wang, R. Wan, S. Wang, T.-Q. Li, and A. Kot, "Domain generalization for medical imaging classification with linear-dependency regularization," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 3115–3126.

[29] R. Gu, J. Zhang, R. Huang, W. Lei, G. Wang, and S. Zhang, "Domain composition and attention for unseen-domain generalizable medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Cham, Switzerland: Springer, 2021, pp. 241–250.

[30] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. Hospedales, "Episodic training for domain generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1446–1455.

[31] C. Li, Q. Qi, X. Ding, Y. Huang, D. Liang, and Y. Yu, "Domain generalization on medical imaging classification using episodic training with task augmentation," 2021, *arXiv:2106.06908*.

[32] J. A. Onofrey *et al.*, "Generalizable multi-site training and testing of deep neural networks using image normalization," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 348–351.

[33] X. Zhao *et al.*, "Robust white matter hyperintensity segmentation on unseen domain," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2021, pp. 1047–1051.

[34] J. He, Z. Deng, and Y. Qiao, "Dynamic multi-scale filters for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3561–3571.

[35] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for RGB-D salient object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 235–252.

[36] J. Zhou, V. Jampani, Z. Pi, Q. Liu, and M.-H. Yang, "Decoupled dynamic filter networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6647–6656.

[37] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," 2021, *arXiv:2102.04906*.

[38] B. Klein, L. Wolf, and Y. Afek, "A dynamic convolutional layer for short range weather prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4840–4848.

[39] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *Proc. 16th Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 282–298.

[40] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "DoDNet: Learning to segment multi-organ and tumors from multiple partially labeled datasets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1195–1204.

[41] N. Bloch *et al.*, "NCI-ISBI 2013 challenge: Automated segmentation of prostate structures," *Cancer Imag. Arch.*, vol. 370, p. 6, Aug. 2015.

[42] G. Lemaître, R. Martí, J. Freixenet, J. C. Vilanova, P. M. Walker, and F. Meriaudeau, "Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review," *Comput. Biol. Med.*, vol. 60, pp. 8–31, May 2015.

[43] G. Litjens *et al.*, "Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge," *Med. Image Anal.*, vol. 18, no. 2, pp. 359–373, 2014.

[44] K. Clark *et al.*, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1045–1057, Dec. 2013.

[45] E. Tsai *et al.* (2020). *Medical Imaging Data Resource Center (MIDRC)—RSNA International COVID-19 Open Radiology Database (RICORD) Release 1a—Chest CT COVID+ (MIDRC-RICORD-1a)*. [Online]. Available: https://wiki.cancerimagingarchive.net/x/DoDTB

[46] J. Sivaswamy *et al.*, "A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis," *JSM Biomed. Imag. Data Papers*, vol. 2, no. 1, p. 1004, Mar. 2015.

[47] F. Fumero, S. Alayon, J. L. Sanchez, J. Sigut, and M. Gonzalez-Hernandez, "RIM-ONE: An open retinal image database for optic nerve evaluation," in *Proc. 24th Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2011, pp. 1–6.

[48] J. I. Orlando *et al.*, "REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Med. Image Anal.*, vol. 59, Jan. 2020, Art. no. 101570.

[49] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[50] C. Chen, K. Hammernik, C. Ouyang, C. Qin, W. Bai, and D. Rueckert, "Cooperative training and latent space data augmentation for robust medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Cham, Switzerland: Springer, 2021, pp. 149–159.

[51] C. Ouyang *et al.*, "Causality-inspired single-source domain generalization for medical image segmentation," 2021, *arXiv:2111.12525*.