

Single-domain Generalization in Medical Image Segmentation via Test-time Adaptation from Shape Dictionary

Quande Liu¹, Cheng Chen¹, Qi Dou¹, Pheng-Ann Heng^{1,2}

¹ Department of Computer Science and Engineering, The Chinese University of Hong Kong

² Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

Abstract

Domain generalization typically requires data from multiple source domains for model learning. However, such strong assumption may not always hold in practice, especially in medical field where the data sharing is highly concerned and sometimes prohibitive due to privacy issue. This paper studies the important yet challenging single domain generalization problem, in which a model is learned under the worst-case scenario with only one source domain to directly generalize to different unseen target domains. We present a novel approach to address this problem in medical image segmentation, which extracts and integrates the semantic shape prior information of segmentation that are invariant across domains and can be well-captured even from single domain data to facilitate segmentation under distribution shifts. Besides, a test-time adaptation strategy with dual-consistency regularization is further devised to promote dynamic incorporation of these shape priors under each unseen domain to improve model generalizability. Extensive experiments on two medical image segmentation tasks demonstrate the consistent improvements of our method across various unseen domains, as well as its superiority over state-of-the-art approaches in addressing domain generalization under the worst-case scenario.

Introduction

Deep networks are notoriously difficult to generalize to new domains due to distribution shift between training and testing data acquired at different situations. Domain generalization (DG) has recently emerged to improve the generalizability of deep networks on unseen domains. Most existing DG methods typically require to learn from multi-source distributions to extract representations that are robust to distribution shifts (Li et al. 2017b). However, the prerequisite to aggregate multi-domain data may not be always achievable in practice. In the field of medical imaging, for example, the data present privacy concerns and are prohibitive to be shared before taking complex procedures including ethical approval and patient consent (Price and Cohen 2019). In this regard, directly generalizing a deep model trained with only single-domain data to any unseen domain is increasingly important and has wider applicability in medical applications.

This problem is *single domain generalization* (SDG), the worst-case scenario of DG with only one source domain

available for model learning. Under this constraint, previous DG methods depending heavily on accessing multi-domain data can hardly work well or lack feasibility in the SDG setting. For instance, the adversarial learning method (Li et al. 2018a) learns domain-invariant representations by regularizing the feature space of multiple data sources, which is infeasible in the case when only one domain is provided for training. Recently, a few studies have started to address SDG problem with the typical solution of adversarial data augmentation (Volpi et al. 2018; Qiao, Zhao, and Peng 2020; Li et al. 2021), which aims to synthesize fictitious samples from single-domain data to potentially simulate the data distributions of unseen domains. However, since it is extremely difficult to anticipate the distribution of test data, these methods are less effective in medical imaging where the domain discrepancies could be large due to complex factors and hardly be simulated from source data (e.g., the domain shift caused by whether using endorectal coil in prostate MRI scanning). Moreover, medical data are usually high-dimensional with tremendous tissue details, making it non-trivial to create high-quality virtual data even after careful model tuning.

Regarding these limitations, we instead aim to address the challenging SDG problem in a simple yet effective way, which is inspired by two key observations in medical image segmentation tasks. First, though the image appearances differ across clinical centers, the semantic shape of segmentation is highly consistent across data sources reflecting the general anatomical structures in medical images. Encouragingly, such shape information can be well-represented even within one single domain, without necessarily aggregating data from multiple clinical sources. Second, though the target distributions of large discrepancies can hardly be simulated during model training, they can instead be explored at test time by considering the given inference sample as a hint. This enables us to mitigate the image distribution bias of single-domain training by adjusting model parameters for each unseen data distribution accordingly at test time.

Based on these insights, we present a new method for addressing SDG named *Test-time Adaptation from Shape Dictionary* (TASD). The core idea is to integrate the general semantic shapes extracted from single-domain data into the segmentation network, and further effectively leverage these prior information through test-time adaptation to generalize the single-domain model to any unseen domains. To this

end, our T ASD first establishes a dictionary learning strategy to extract a set of explicit shape priors from the single source domain. A regression branch is then embedded onto the network, which produces weighting coefficients to integrate these dictionary items, so that the segmentation could be derived by jointly recognizing the sample features as well as the pre-collected shape priors to facilitate generalization. Next, when deploying the model to new domains, we have devised a dual-consistency regularization mechanism to adapt the model parameters at test time, which jointly enforces the consistency of shape coefficients and segmentation predictions under different input perturbations applied to the inference data. Such adaptation strategy allows to dynamically update the model parameters especially for the regression branch at test time, which enables the model to adaptively utilize the shape prior information under any unseen data distributions to improve model generalizability.

Our main contributions are highlighted as follows:

- We present a novel approach to address the challenging single domain generalization problem for medical image segmentation, by explicitly exploiting the general semantic shape priors that are extractable from single-domain data and are generalizable across domains to assist domain generalization under the worst-case scenario.
- We also devise an effective test-time adaptation strategy with dual-consistency regularization, which allows to adaptively utilize the shape prior information at any unseen data distributions to improve model generalizability.
- We conduct extensive experiments on two typical medical image segmentation tasks, i.e., prostate MRI and retinal fundus image segmentation. With only single-domain training, our method significantly improves generalization performance on many unseen domains, outperforming the state-of-the-arts. Code will be released publicly.

Related Work

Domain generalization aims to generalize models to completely unseen domains without accessing target data during training (Chattopadhyay, Balaji, and Hoffman 2020; Du et al. 2020; Hoffer et al. 2020; Liu et al. 2021; Seo et al. 2020; Yue et al. 2019; Fan et al. 2021; Dou et al. 2021). Previous progress on DG is mainly made by enabling models to learn from multi-source domain data to reduce the model bias. A plenty of methods aim to extract domain-invariant representations by aligning the feature space of multiple data sources (Muandet, Balduzzi, and Schölkopf 2013; Ghifary et al. 2015; Hsu, Lv, and Kira 2017; Motiian et al. 2017; Li et al. 2018a,b). Some other studies have introduced self-supervised learning (Huang et al. 2020) or model-agnostic meta-learning (Li et al. 2017a; Dou et al. 2019; Balaji, Sankaranarayanan, and Chellappa 2018; Li et al. 2019) to regularize the model training with multiple domains to learn generalizable representations. Recently, the worst-case scenario with only one source domain available for model training has been attempted in (Qiao, Zhao, and Peng 2020; Volpi et al. 2018; Wang et al. 2021b; Li et al. 2021) with adversarial domain augmentation. Typically, (Qiao, Zhao, and Peng

2020) propose a method which incorporates meta-learning and Wasserstein auto-encoder to facilitate the learning efficiency from generated data. Considering the large domain discrepancies and the difficulty in data generation for medical images, we instead tackle the challenging SDG problem from a new perspective by leveraging the domain-invariant shape information of medical anatomies and promoting effective test-time utilization of these learned shape priors.

Dictionary learning seeks to find a set of basic elements to compose a dictionary such that a given input can be well represented by a sparse linear combination of these learned elements. Successful applications of dictionary learning have emerged in various image recognition tasks (Zhang et al. 2012; Heimann and Meinzer 2009). For example, dictionary learning is employed in (Vu and Monga 2017) to learn a set of common patterns and class-specific patterns from different objects for image classification. (Cao et al. 2015) present semantic label dictionary which explicitly fuses the label information into dictionary representation and explores the semantic correlations between co-occurrence labels for multi-label image annotation. In medical applications, dictionary learning is conducted in (Tong, Wolz et al. 2015) to find a set of most representative atlases, which are then applied for label propagation to solve a multi-organ segmentation task. In our work, we employ dictionary learning to explicitly extract the shape priors of medical anatomies from single-domain data, and more importantly, we integrate the learned shape dictionary into deep models to facilitate the generalization on unseen domains. To our best knowledge, this is the first work to leverage dictionary learning to solve model generalization problem.

Test-time learning is recently proposed to utilize the distribution information from the target data presented at test time to quickly adapt models with a few gradient steps (Chen et al. 2021; Pandey et al. 2021). The main differences of prior works lie in how to devise the objective for driving test-time learning and which part of network parameters to be updated at test time. For instance, the pioneer work (Sun et al. 2020) exploits the test data via an auxiliary branch with self-supervision to adapt the encoder of a classification model. Later on, Wang et al. (2021a) penalize the entropy of model predictions on test data to adapt the batch normalization layer for generalization purpose. For test-time adaptation on medical image segmentation, Karani et al. (2021) introduce a denoising autoencoder into test-time learning to adapt an image normalization module. This method requires iterative refinement of model predictions and takes nearly 12 minutes for the adaptation on one medical volume data, which therefore cannot satisfy the real-time adaptation requirement in clinical scenarios. Compared with these methods, our test-time update is tightly-coupled with the learned generalizable shape dictionary, and is driven by the lightweight dual-consistency regularization mechanism which enables the model to dynamically utilize the shape priors according to various data distributions at test time.

Methodology

An overview of our method *Test-time Adaptation from Shape Dictionary* is shown in Fig. 1. This section describes the two

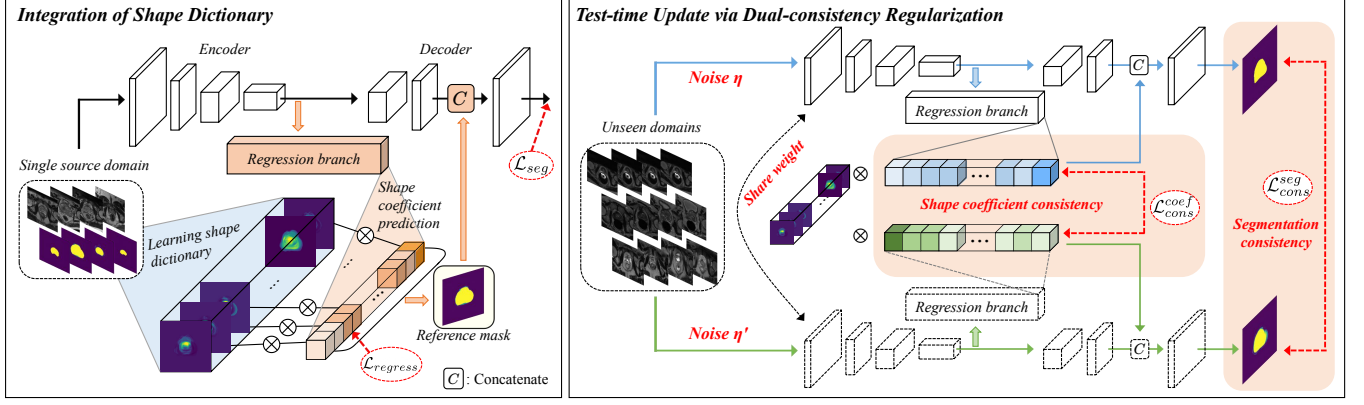


Figure 1: Overview of our proposed *Test-time Adaptation from Shape Dictionary* (TASD). During training, we establish dictionary learning to extract the explicit shape priors from single domain data and integrate these prior information into the network by introducing a regression branch to produce shape coefficients for combining these dictionary items. When testing on new domains, we dynamically adapt the model parameters especially on the regression branch to adaptively utilize the prior information to improve generalization, which is driven by dual-consistency regularization mechanism which jointly regularizes the consistency of shape coefficients and segmentation predictions under different perturbations (η, η') onto the inference data.

main components of TASD followed by its technical details of source domain training and test-time update.

Integrating Shape Dictionary for SDG

Construction of shape dictionary: The semantic shapes of segmentation reflect the underlying anatomical structures in medical images, thus are independent of the scanner effects from observed domains. In light of this, we expect to harness this general shape information, which can be well-captured even within single source domain, to boost model generalization in SDG. Motivated by the nature of dictionary learning to learn representative elements, we employ it to explicitly extract the representative shape priors from single-domain data, and further effectively integrate the shape information into the task model to provide references for the segmentation at unseen domains.

Specifically, dictionary learning is to find a set of basic elements (called dictionary), such that a given input can be approximated as their sparse linear combinations. In our case, denote $\mathbf{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ as the single source domain data containing N pairs of data and labels. We intend to find a dictionary \mathbf{D} consisting of K explicit shape templates, i.e., $\mathbf{D} = \{\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_K\}$, that is able to represent the diverse segmentation masks $\{\mathbf{y}_i\}_{i=1}^N$ in the task. To ensure that the shape dictionary learned from one domain is representative enough to be extended to unseen domains, two principles need to be considered for its construction: (1) $K \ll N$, the number of shape templates should be much smaller than the number of masks in the single source data (e.g., $K = 48$ for $N = 400$), to reduce overfitting of the learned dictionary, (2) the mask \mathbf{y}_i should be represented with as few elements as possible (i.e., the coefficients for linear combination should be sparse), to maximize the representation power of each template in the shape dictionary. For simplicity, we consider a binary segmentation task (i.e., $\mathbf{y}_i \in \mathbb{R}^{H \times W}$), while the extension to multi-class segmentation is straight-

forward by constructing a shape dictionary for each class in the same way. Then the generation of dictionary is equivalent to solving the optimization problem with respect to the shape dictionary $\mathbf{D} \in \mathbb{R}^{K \times H \times W}$ and the coefficients $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_N]$ ($\alpha_i \in \mathbb{R}^K$ denotes the learned coefficients of dictionary for each \mathbf{y}_i) as:

$$\arg \min_{\mathbf{D}, \boldsymbol{\alpha}} \sum_{i=1}^N (\|\mathbf{y}_i - \sum_{j=1}^K \mathbf{d}_j \alpha_{ij}\|_2^2 + \lambda \|\boldsymbol{\alpha}_i\|_1), \quad (1)$$

where λ is a balancing parameter to regularize the learned coefficients. In our implementation, we fixed λ as $1.2/\sqrt{m}$ (m denotes the input resolution, i.e., $H \times W$) as this is a general solution in dictionary learning to yield sparse coefficients. The optimization problem in Eq. 1 can be solved using the well-established online dictionary learning algorithm (Mairal et al. 2009), which iteratively draws each image mask \mathbf{y}_i to find its sparse coefficients with least-angle regression and update the dictionary \mathbf{D} with block-coordinate descent. Meanwhile, this learning algorithm is parameter-free hence does not need parameter-tuning. We show in experiments that the shape dictionary learned in this way is representative and can well generalize to unseen domains.

Integration of shape dictionary into deep network: After obtaining the learned shape dictionary, we further incorporate it into the deep model, aiming to enable the segmentation to be performed with references from these shape priors for boosting generalization. To this end, we embed a regression branch onto the network (see the left part of Fig. 1), which takes the encoded features of a given sample as input, and output the shape coefficients $\hat{\boldsymbol{\alpha}}$ for combining the shape priors according to the sample features. Specifically, the pre-extracted shape priors in \mathbf{D} are linearly combined according to the shape coefficients $\hat{\boldsymbol{\alpha}}$ to generate a reference mask:

$$\mathbf{M} = \sum_{j=1}^K \mathbf{d}_j \hat{\alpha}_{ij} \quad (2)$$

which is then concatenated with network features before the last convolutional block of decoder for further refinement and generate the final segmentation mask. We did not choose to insert the reference mask at early layers since this requires downsampling the reference mask, which would destroy the shape prior information to some extent. With the help of the embedded regression branch, the model predictions can be derived by jointly exploring the sample features as well as the shape priors extracted from single-domain data to improve model generalization.

To provide direct supervision for the regression branch to improve the quality of reference mask, we fetch the coefficients α optimized from Eq. 1 (obtained during dictionary learning) as the ground truth of $\hat{\alpha}$. Since α is regularized to be sparse during dictionary learning, we adopt the cosine similarity to minimize the distance between $\hat{\alpha}$ and α instead of the L1 or L2 loss, which is expressed as:

$$\mathcal{L}_{regress}(\mathbf{x}_i) = 1 - \frac{\hat{\alpha}_i \cdot \alpha_i}{\max(\|\hat{\alpha}_i\|_2 \cdot \|\alpha_i\|_2, \epsilon)}, \quad (3)$$

where ϵ is a small value set as 1e-8 to avoid division by zero.

Dual-consistency Regularized Test-time Update

With embedding an regression branch onto the network, the shape priors have been integrated to assist the segmentation problem. However, a potential issue is that the model's ability to utilize these prior information (i.e., parameters of the regression branch and original network) are learned under the narrow single-domain data distributions. When generalizing to unseen domains with distribution shifts, the reliability of generated shape coefficients cannot be guaranteed, which would limit the shape priors to be effectively utilized. Given the observation that the inference sample presented at test time can give us a hint about its distribution information, we expect to dynamically adapt the model parameters at test time by exploring this hint, so that the shape priors can be effectively utilized under any unseen data distributions to improve generalization performance.

Crucially, the test-time update in our case should be tightly-coupled with the regression branch to ensure the correctness of shape coefficients under domain shifts. To this end, we formulate the adaptation as a consistency regularization mechanism, which can be flexibly imposed on both original network as well as the embedded regression branch. Specifically, as shown in the right part of Fig. 1, given an inference data \mathbf{x}^t at test time, we first add different perturbations η and η' (e.g., adding Gaussian noise to input or dropout to network parameters) onto this sample, and then regularize the model to generate consistent predictions for the two perturbed samples to explore the inherent information of this testing data. Importantly, to fully adapt the parameters of regression branch apart from the original network, we impose the consistency regularization on both segmentation predictions and shape coefficients predictions. The intuition behind is that different input perturbations should not change the semantic anatomies contained in a medical image, hence the final segmentation mask of an inference sample as well as how our model leverages the

pre-collected shape priors (i.e., shape coefficients) to compose the mask should not be changed. Formally, these two consistency regularization terms can be formulated as:

$$\begin{aligned} \mathcal{L}_{cons}^{seg}(\mathbf{x}^t) &= \|f(\mathbf{x}^t, \eta') - f(\mathbf{x}^t, \eta)\|_2^2, \\ \mathcal{L}_{cons}^{coef}(\mathbf{x}^t) &= 1 - \frac{\hat{\alpha}^t(\eta') \cdot \hat{\alpha}^t(\eta)}{\max(\|\hat{\alpha}^t(\eta')\|_2 \cdot \|\hat{\alpha}^t(\eta)\|_2, \epsilon)}, \end{aligned} \quad (4)$$

where $\mathcal{L}_{cons}^{seg}(\mathbf{x}^t)$ is the consistency regularization on segmentation results with mean square error loss, in which $f(\cdot)$ denotes the network, η and η' are different perturbations applied to the same testing data \mathbf{x}^t ; and $\mathcal{L}_{cons}^{coef}(\mathbf{x}^t)$ denotes the consistency regularization on shape coefficients with cosine distance loss, in which $\hat{\alpha}^t(\cdot)$ is the shape coefficient predictions for \mathbf{x}^t under perturbation (\cdot) . Our dual-consistency regularization objective can then be expressed as:

$$\mathcal{L}_{cons}^{dual}(\mathbf{x}^t) = \mathcal{L}_{cons}^{seg}(\mathbf{x}^t) + \mathcal{L}_{cons}^{coef}(\mathbf{x}^t), \quad (5)$$

At test time, we only perform one step of gradient descent by optimizing Eq. 5 to adapt the parameters of source model given each inference data. Based on our experiments, such adaptation strategy only requires in average 56ms to process a 2D medical image and 0.91s for a medical volume data, hence can be performed efficiently. Meanwhile, we only adapt the feature normalization parameters, i.e., the parameters of the batch normalization layers in the source model, which enables the model to adaptively normalize the features according to the data distributions of each unseen domain. This yields more stable results in our experiments compared with adapting all network parameters, which could be explained by that the deep network is overly parameterized hence adapting the whole model is prone to affect its original discriminability. It is worthy to note that the adaptation strategy in our method is performed in an online manner, i.e., when receiving test sample \mathbf{x}_j^t at time point j , the model state is initialized with the parameters updated from the previous sample \mathbf{x}_{j-1}^t instead of the original source model. This enables the model to fully utilize the distribution information from all previous samples to boost generalization, and is also consistent with the clinical scenario where testing samples usually arrive sequentially.

Training Objectives and Strategies

In our framework, we first learned the shape dictionary from single-domain data by solving Eq. 1 with online dictionary learning algorithm (Mairal et al. 2009). Then, we integrated the shape dictionary that contains the shape prior information into segmentation network, and the overall objective to train the network at single source domain is expressed as:

$$\mathcal{L}(\mathbf{x}_i) = \mathcal{L}_{seg}(\mathbf{x}_i) + \beta \mathcal{L}_{regress}(\mathbf{x}_i), \quad (6)$$

which is composed of the segmentation loss and the shape coefficient regression loss with a weighting parameter β . We empirically set β as 5 to balance the scale of the two loss terms. Once the training is done, we can generalize the model to unseen domains by dynamically adapting the model parameters as described above.

Experiment

We extensively evaluate our approach on two typical medical image segmentation tasks: 1) prostate segmentation on T2-weighted MRI, and 2) optic disc and cup segmentation on retinal fundus images. Results of comparison with state-of-the-arts and ablation study are presented in the following.

Datasets and Evaluation Metrics

Prostate MRI segmentation: We collect prostate T2-weighted MRIs from six different clinical centers out of three public datasets, including NCI-ISBI13 (Bloch et al. 2015), I2CVB (Lemaître et al. 2015) and PROMISE12 (Litjens et al. 2014) datasets. These data are uniformly resized to 384×384 in axial plane, and each data volume is normalized to have zero mean and unit variance in intensity values. Due to the large variance on the slice thickness of data from different clinical sites, we employ a 2D network as backbone.

Fundus image segmentation: We employ retinal fundus images from four medical institutions out of three public datasets, including REFUGE (Orlando et al. 2020), Drishti-GS (Sivaswamy et al. 2015) and RIM-ONE-r3 (Fumero et al. 2011) datasets. For this task, we follow the fundus image pre-processing pipeline of (Wang et al. 2020) to first train a simple U-Net to detect and crop ROIs around optic disc with the size of 800×800 , which are then resize to 384×384 as network inputs. Since the ROI size of 800×800 is about twice of the optic disc region, the location deviation is relaxed and we found that all optic disc and cup regions are covered by the cropped ROI even under data distribution shifts. We average the performance of optic disc and cup segmentation as the task performance.

In both tasks, data acquired from different clinical centers present discrepancies due to different imaging conditions. The representative cases and sample numbers of each dataset are shown in Fig. 2. For the evaluation, we employ two commonly-used metrics to quantitatively evaluate the segmentation results, including the Dice coefficient [%] (the higher the better) and Hausdorff Distance (HD) [pixel] (the lower the better), which evaluate the segmentation results in terms of the whole object and surface region, respectively.

Implementation Details

In our implementation, the element number K in the shape dictionary is set as 48, and we will investigate its effect in the ablation study. For input perturbations, two dropout layers with dropout rate 0.5 are added before the bottleneck layer and the last convolution layer respectively, and Gaussian noise with magnitude 0.1 is added onto the input images. The step size for test-time update is $1e-3$. We employ an adapted Mix-residual-UNet (Liu et al. 2020) as segmentation backbone. The standard data augmentation techniques are used to avoid overfitting, including random rotation and flipping (horizontal and vertical). The model is trained using Adam optimizer with momentum of 0.9 and 0.99, and the learning rate is initialized as $1e-3$. We totally train 100 epochs on the single source domain as the network has converged, with batch size set as 5. The framework is implemented with Pytorch using one NVIDIA TitanXp GPU.

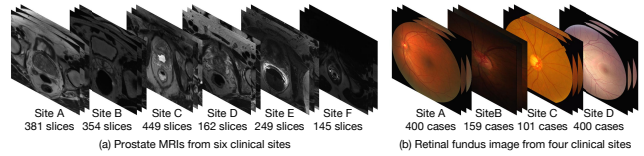


Figure 2: Representative cases and case number of each data source for prostate MRI and retinal fundus images.

Results of Single Domain Generalization

Experimental settings: To fulfill SDG setting, for prostate MRI segmentation, we follow (Karani et al. 2021) to use a single domain dataset NCI-ISBI13 as the source domain (Site A) and take the other five data sources as different unseen target sites (Site B-F). For fundus image segmentation, we adopt the popular benchmark that is the training set of REFUGE as the single source domain (Site A), and the other three data sources as different target sites (Site B-D).

We compare our TASD with recent state-of-the-art methods in SDG and test-time learning, as well as popular DG methods that are not completely restricted to multiple source domains and can be applied in the SDG scenario, including: **M-ADA (CVPR 2020)** (Qiao, Zhao, and Peng 2020): a SDG method with adversarial domain augmentation, which develops Wasserstein auto-encoder to create virtual data. **TTT (ICML 2020)** (Sun et al. 2020) a test-time learning method for generalization under distribution shifts with self-supervision of rotation prediction. **TTST (MIA 2021)** (Karani et al. 2021) a test-time learning method for generalization in medical image segmentation with self-training from a denoising autoencoder. **Tent (ICLR 2021)** (Wang et al. 2021a): a test-time adaptation method by minimizing the entropy of model predictions. **BigAug (TMI 2020)** (Zhang et al. 2020): a DG method on medical image segmentation that adopts extensive data transformations to promote general representation learning. **JiGen (CVPR 2019)** (Carlucci et al. 2019): a DG method with self-supervised learning that learns general representations by solving jigsaw puzzles. In SDG scenario, the **Baseline** setting denotes learning a model on the single source domain without using any generalization technique, which is different from general DG scenario where the baseline model is obtained by training with multiple source datasets (Liu, Dou, and Heng 2020; Wang et al. 2020).

Comparison results: Table 1 shows the results on prostate MRI segmentation. We see that different methods can generally improve the generalization performance over baseline. The M-ADA approach dedicatedly designed for the SDG setting performs better than the DG methods with self supervision (JiGen) or traditional data augmentation (BigAug). Compared with M-ADA, our TASD presents clear Dice increase on all unseen sites, with improvement of 2.9% and 8.7 on the average performance of Dice and HD respectively. This can be attributed to our explicit integration of shape priors and effective utilization of the prior knowledge at test time to improve generalization. For the test-time learning methods, TTST outperforms TTT and Tent in our segmentation problem, owing to their self-training strat-

Table 1: Quantitative comparison of single domain generalization results on prostate MRI segmentation between different methods.

Unseen Site	B	C	D	E	F	Avg.	B	C	D	E	F	Avg.
	Dice Coefficient (Dice, mean \pm std) \uparrow						Hausdorff Distance (HD, mean \pm std) \downarrow					
Baseline	83.8 \pm 5.3	73.3 \pm 11.1	72.6 \pm 7.0	65.5 \pm 29.5	78.7 \pm 7.8	74.8	40.9 \pm 34.9	59.0 \pm 29.4	59.5 \pm 21.2	61.2 \pm 42.6	37.2 \pm 20.9	51.6
M-ADA (Qiao, Zhao, and Peng 2020)	86.2 \pm 4.4	74.7 \pm 9.1	80.9 \pm 4.9	69.7 \pm 12.2	79.5 \pm 9.3	78.2	19.1\pm21.1	46.1 \pm 28.1	53.9 \pm 19.3	54.2 \pm 19.6	31.9 \pm 26.7	41.0
TTT (Sun et al. 2020)	83.5 \pm 5.9	73.1 \pm 17.5	75.3 \pm 7.8	67.5 \pm 11.1	81.5 \pm 5.9	76.2	26.4 \pm 22.1	55.4 \pm 22.3	54.8 \pm 25.5	53.0 \pm 22.2	21.8 \pm 19.2	42.3
TTST (Karani et al. 2021)	86.0 \pm 3.7	74.8 \pm 10.5	81.0 \pm 3.9	74.0 \pm 8.4	80.9 \pm 9.2	79.3	20.5 \pm 20.7	47.5 \pm 28.1	41.4 \pm 19.7	51.4 \pm 26.1	34.5 \pm 25.5	39.1
Tent (Wang et al. 2021a)	84.5 \pm 4.7	74.2 \pm 13.9	76.4 \pm 8.1	67.1 \pm 10.1	80.1 \pm 9.6	76.5	27.2 \pm 24.7	50.3 \pm 22.7	45.7 \pm 23.5	49.6 \pm 30.9	29.8 \pm 20.1	40.5
BigAug (Zhang et al. 2020)	84.2 \pm 5.0	73.9 \pm 14.1	73.3 \pm 7.7	74.7 \pm 9.7	79.0 \pm 6.8	77.0	35.9 \pm 26.3	49.1 \pm 20.7	53.8 \pm 22.0	44.5 \pm 18.7	28.9 \pm 14.4	42.4
JiGen (Carlucci et al. 2019)	83.2 \pm 6.1	70.8 \pm 14.7	74.0 \pm 7.9	71.5 \pm 10.2	80.3 \pm 6.2	75.9	29.3 \pm 23.2	64.5 \pm 23.3	50.4 \pm 25.9	50.6 \pm 26.0	24.3 \pm 10.7	43.8
Baseline + ISD	85.4 \pm 5.1	74.0 \pm 13.0	79.0 \pm 4.4	72.7 \pm 9.2	80.8 \pm 10.0	78.4	19.4 \pm 21.3	54.8 \pm 24.8	44.2 \pm 21.7	48.9 \pm 23.6	34.9 \pm 26.4	40.4
Baseline + TTA	84.7 \pm 8.2	75.2 \pm 8.2	74.8 \pm 8.2	67.2 \pm 8.2	79.7 \pm 8.2	76.3	25.7 \pm 23.4	42.0 \pm 20.2	46.1 \pm 21.9	47.9 \pm 37.4	30.3 \pm 20.6	38.4
TASD (Ours)	87.1\pm2.5	76.4\pm6.1	82.5\pm5.2	76.0\pm6.6	83.2\pm6.7	81.1	19.3 \pm 21.3	39.1\pm17.5	38.7\pm12.2	43.4\pm14.2	21.0\pm17.5	32.3

Table 2: Quantitative comparison of single domain generalization results on fundus image segmentation between different methods.

Unseen Site	B	C	D	Avg.	B	C	D	Avg.
	Dice Coefficient (Dice, mean \pm std) \uparrow				Hausdorff Distance (HD, mean \pm std) \downarrow			
Baseline	83.2 \pm 8.2	76.0 \pm 14.8	88.8 \pm 5.7	82.7	27.4 \pm 15.3	36.8 \pm 23.8	21.5 \pm 10.3	28.6
M-ADA (Qiao, Zhao, and Peng 2020)	85.9 \pm 9.3	77.4 \pm 13.7	90.6 \pm 4.8	84.6	22.7 \pm 11.2	30.4 \pm 17.1	13.8 \pm 7.5	22.3
TTT (Sun et al. 2020)	84.6 \pm 5.7	77.3 \pm 11.9	89.0 \pm 4.1	83.7	22.2 \pm 8.6	31.4 \pm 19.1	17.4 \pm 7.1	23.7
TTST (Karani et al. 2021)	85.5 \pm 9.0	76.5 \pm 13.4	91.0 \pm 4.7	84.3	22.0 \pm 10.2	33.6 \pm 19.3	12.1\pm8.1	22.6
Tent (Wang et al. 2021a)	85.1 \pm 8.1	77.1 \pm 13.0	89.2 \pm 4.9	83.8	23.1 \pm 11.7	35.2 \pm 20.9	17.8 \pm 8.5	25.4
BigAug (Zhang et al. 2020)	84.7 \pm 7.5	78.0 \pm 13.5	90.7 \pm 4.6	84.5	27.1 \pm 13.2	30.3 \pm 21.6	14.8 \pm 7.9	24.1
JiGen (Carlucci et al. 2019)	84.5 \pm 5.8	77.5 \pm 12.0	88.5 \pm 4.1	83.5	23.4 \pm 9.2	34.3 \pm 19.4	20.4 \pm 10.6	26.0
Baseline + ISD	85.7 \pm 8.0	78.1 \pm 12.9	89.5 \pm 5.9	84.4	22.7 \pm 10.0	29.6 \pm 17.9	15.2 \pm 9.2	22.5
Baseline + TTA	83.6 \pm 7.6	77.2 \pm 13.8	89.6 \pm 5.6	83.5	26.0 \pm 13.9	34.1 \pm 22.6	15.8 \pm 10.6	25.3
TASD (Ours)	87.6\pm8.0	78.5\pm12.6	91.3\pm4.2	85.8	19.8\pm9.5	29.4\pm18.0	12.3 \pm 6.2	20.5

egy to gradually refine the segmentation masks. Our TASD performs clearly better than TTST in the average performance on both two metrics. This endorses the benefits of the general shape priors integrated to the task model, which provides strong reference for the test-time update to facilitate the segmentation at unseen domains. Overall, TASD achieves the best average performance and consistently outperforms the baseline on all unseen sites by a large margin.

Table 2 shows the results on fundus image segmentation. Similarly, all comparison methods can more or less improve the average generalization performance, and the other observations drawn in prostate MRI segmentation also hold here. Our TASD beats all comparison methods in the average performance, with clear improvements over baseline on all three unseen sites. This again confirms the superiority of our method to address SDG for medical image segmentation. Fig. 3 further shows the segmentation results with two example unseen cases for each task. It is observed that TASD can accurately preserve the anatomy shapes at unknown distributions, whereas other methods sometimes fail to do so.

Ablation Study

We conduct experiments to investigate several key points in our method: 1) the contribution of the two components of integration of shape dictionary (ISD) and test-time adaptation (TTA) in our method; 2) the representation power of the shape dictionary; 3) the effect of dictionary size and number of masks to learn the shape dictionary; 4) the effect of the two consistency regularization terms at test-time adaptation.

Contribution of the two components. As shown in Ta-

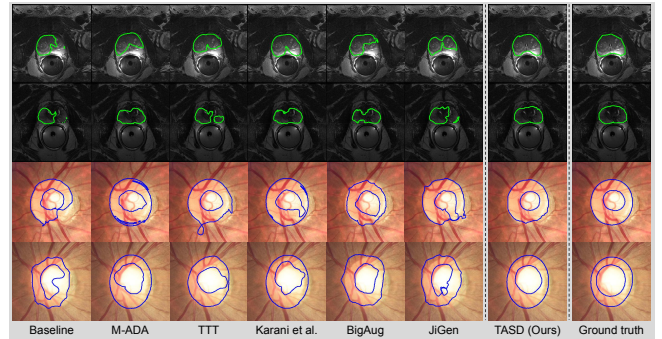


Figure 3: Qualitative comparison for different methods under SDG scenario, with first two rows for prostate MRI segmentation and last two rows for fundus image segmentation.

ble 1 and Table 2, integrating the shape dictionary over baseline (Baseline + ISD) yields consistent improvements on different unseen sites, which demonstrates the benefits of utilizing the general shape information to enhance model generalization under the worst single domain training scenario. Further adapting the model at test time completes our TASD method, which increases the average Dice over Baseline + ISD by 2.7% and 1.4% for the two tasks. Notably, such improvements are higher than the performance gains when deploying test-time adaptation on the baseline model (i.e., Baseline + TTA vs. Baseline), which is 1.5% and 0.8% for the two tasks. This observation reflect that our test-time adaptation mechanism gains additional capacity when de-

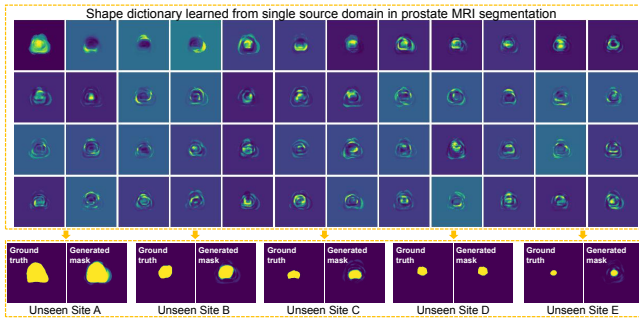


Figure 4: Visualization of the shape dictionary learned from single source domain with $K=48$, as well as the mask derived from the dictionary for a representative sample at each unseen site (with ground truth displayed for reference).

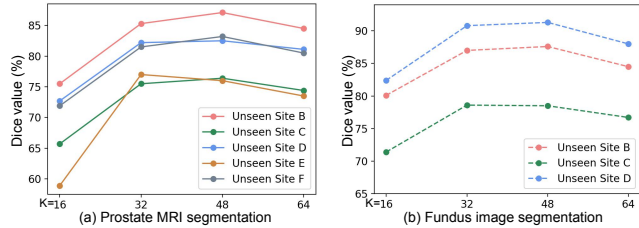


Figure 5: Effect of different dictionary size K on the generalization performance.

played over the network with integrated shape dictionary, and enables the model to adaptively utilize the shape priors from single-domain data to improve model generalizability.

Representation power of the learned shape dictionary.

Fig. 4 displays the learned shape dictionary from single-domain data with dictionary size $K=48$ and the mask derived from the dictionary for representative samples of each unseen site. We can see that the masks derived from pre-collected shape priors for different unseen-site samples are highly consistent with their ground truth masks with various sizes. We also directly compute the accuracy of reference mask \bar{M} (cf. Eq. 2) generated from shape dictionary, observing that it attains 77.1% and 83.9% Dice for the averaged generalization performance in prostate and fundus tasks, which is clearly higher than the baseline model. These results indicate that the shape dictionary learned from single-domain data is capable of representing the diverse masks at each unseen site, and affirm our motivation to utilize the general shape priors information to address the challenging single domain generalization problem.

Effect of dictionary size K and number of ground truth masks N . The choice of element number K in dictionary is important in our method, which affects the representation power of the extracted shape priors. Intuitively, less elements in the shape dictionary (i.e., smaller K) might be incapable to represent the diverse segmentation masks, while too more elements (i.e., larger K) could lead to overfitting of the dictionary and make each shape template less representative to generalize to unseen domains. To investigate the suitable choice of K , we repeat the experiment of

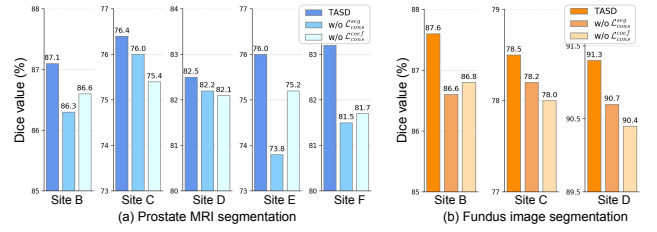


Figure 6: Ablation analysis of the dual-consistency regularization mechanism at test-time adaptation.

TASD by varying $K \in \{16, 32, 48, 64\}$. As shown in Fig. 5, the models with middle-level dictionary size ($K=32$ or 48) perform better than the model with smaller ($K=16$) or larger dictionary ($K=64$) on both tasks. These results confirm our analysis above, and we finally adopt $K=48$ in our method. Furthermore, we analyze the effect of N which is the number of ground truth masks used to learn the shape dictionary. Specifically, we randomly sample a certain proportion from all masks of single-domain data, observing that the Dice performances decrease by 0.7% and 1.6% in prostate segmentation, 0.4% and 0.9% in fundus image segmentation tasks when sampling 50% and 25% from the total ground truth masks. This reflects that a larger N is helpful to learn more general shape priors and therefore can better improve the generalization performance under data distribution shifts.

Effect of two consistency regularization terms. To let the model adaptively exploit the integrated shape priors, we have designed a dual-consistency regularization mechanism which imposes the consistency of both segmentation and shape coefficient predictions under input perturbations. As listed in Fig. 6, removing the regularization from either place (w/o $\mathcal{L}_{cons}^{coef}$ or w/o \mathcal{L}_{cons}^{seg}) can lead to performance drop compared with the full TASD model. This demonstrates the indispensable roles of the two regularization terms in our method. Specifically, the $\mathcal{L}_{cons}^{coef}$ allows the model to better utilize the shape prior knowledge at test time, while the \mathcal{L}_{cons}^{seg} helps to better adapt the network parameters and refine the generated reference masks.

Conclusion

We present a new method with promising performance achieved on the challenging SDG problem in medical image segmentation. The idea is to integrate the semantic shape priors that are generalizable across domains and can be well captured from single-domain data to assist segmentation, and further effectively exploit these prior information via test-time adaptation driven by dual-consistency regularization to promote the model generalization at any unseen data distributions. The integration of shape priors extracted with dictionary learning is extendable to address other segmentation problems such as multi-modality learning, and evaluating its benefit is appealing as our future work. Besides, the proposed test-time learning method with consistency regularization mechanism is applicable to other scenarios.

Acknowledgement

This work was supported by Key-Area Research and Development Program of Guangdong Province, China under Grant 2020B010165004, Hong Kong RGC TRS Project No. T42-409/18-R, Hong Kong Innovation and Technology Fund Project No. GHP/110/19SZ, and National Natural Science Foundation of China with Project No. U1813204.

References

- Balaji, Y.; Sankaranarayanan, S.; and Chellappa, R. 2018. Metareg: Towards domain generalization using meta-regularization. In *Advances in Neural Information Processing Systems*, 998–1008.
- Bloch, N.; et al. 2015. NCI-ISBI 2013 challenge: automated segmentation of prostate structures. *The Cancer Imaging Archive*, 370.
- Cao, X.; Zhang, H.; Guo, X.; et al. 2015. Sled: Semantic label embedding dictionary representation for multilabel image annotation. *IEEE Transactions on Image Processing*, 24(9): 2746–2759.
- Carlucci, F. M.; D’Innocente, A.; Bucci, S.; Caputo, B.; and Tommasi, T. 2019. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2229–2238.
- Chattopadhyay, P.; Balaji, Y.; and Hoffman, J. 2020. Learning to balance specificity and invariance for in and out of domain generalization. *arXiv preprint arXiv:2008.12839*.
- Chen, C.; Liu, Q.; Jin, Y.; Dou, Q.; and Heng, P.-A. 2021. Source-Free Domain Adaptive Fundus Image Segmentation with Denoised Pseudo-Labeling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 225–235. Springer.
- Dou, Q.; de Castro, D. C.; Kamnitsas, K.; and Glocker, B. 2019. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems*, 6450–6461.
- Dou, Q.; So, T. Y.; Jiang, M.; Liu, Q.; Vardhanabhuti, V.; Kaissis, G.; Li, Z.; Si, W.; Lee, H. H.; Yu, K.; et al. 2021. Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study. *NPJ digital medicine*, 4(1): 1–11.
- Du, Y.; Xu, J.; Xiong, H.; Qiu, Q.; Zhen, X.; Snoek, C. G.; and Shao, L. 2020. Learning to learn with variational information bottleneck for domain generalization. *arXiv preprint arXiv:2007.07645*.
- Fan, X.; Wang, Q.; Ke, J.; Yang, F.; Gong, B.; and Zhou, M. 2021. Adversarially Adaptive Normalization for Single Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8208–8217.
- Fumero, F.; et al. 2011. RIM-ONE: An open retinal image database for optic nerve evaluation. In *2011 24th international symposium on computer-based medical systems (CBMS)*, 1–6. IEEE.
- Ghifary, M.; Bastiaan Kleijn, W.; Zhang, M.; and Balduzzi, D. 2015. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, 2551–2559.
- Heimann, T.; and Meinzer, H.-P. 2009. Statistical shape models for 3D medical image segmentation: a review. *Medical image analysis*, 13(4): 543–563.
- Hoffer, E.; Ben-Nun, T.; Hubara, I.; Giladi, N.; Hoefler, T.; and Soudry, D. 2020. Augment Your Batch: Improving Generalization Through Instance Repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8129–8138.
- Hsu, Y.-C.; Lv, Z.; and Kira, Z. 2017. Learning to cluster in order to transfer across domains and tasks. *arXiv preprint arXiv:1711.10125*.
- Huang, Z.; Wang, H.; Xing, E. P.; and Huang, D. 2020. Self-Challenging Improves Cross-Domain Generalization. In *ECCV*.
- Karani, N.; Erdil, E.; Chaitanya, K.; and Konukoglu, E. 2021. Test-time adaptable neural networks for robust medical image segmentation. *Medical Image Analysis*, 68: 101907.
- Lemaître, G.; et al. 2015. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review. *Computers in biology and medicine*, 60: 8–31.
- Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2017a. Learning to generalize: Meta-learning for domain generalization. *arXiv preprint arXiv:1710.03463*.
- Li, D.; Yang, Y.; Song, Y.-Z.; et al. 2017b. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, 5542–5550.
- Li, H.; Jialin Pan, S.; Wang, S.; and Kot, A. C. 2018a. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5400–5409.
- Li, L.; Gao, K.; Cao, J.; Huang, Z.; Weng, Y.; Mi, X.; Yu, Z.; Li, X.; and Xia, B. 2021. 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.; Tian, X.; Gong, M.; Liu, Y.; Liu, T.; Zhang, K.; and Tao, D. 2018b. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 624–639.
- Li, Y.; Yang, Y.; Zhou, W.; and Hospedales, T. M. 2019. Feature-Critic Networks for Heterogeneous Domain Generalization. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, 3915–3924. PMLR.
- Litjens, G.; et al. 2014. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. *Medical image analysis*, 18(2): 359–373.
- Liu, Q.; Chen, C.; Qin, J.; Dou, Q.; and Heng, P.-A. 2021. Feddg: Federated domain generalization on medical image

- segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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. Springer.
- Liu, Q.; Dou, Q.; Yu, L.; and Heng, P. A. 2020. MS-Net: multi-site network for improving prostate segmentation with heterogeneous MRI data. *IEEE transactions on medical imaging*, 39(9): 2713–2724.
- Mairal, J.; Bach, F.; Ponce, J.; and Sapiro, G. 2009. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, 689–696.
- Motiian, S.; Piccirilli, M.; Adjeroh, D. A.; and Doretto, G. 2017. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 5715–5725.
- Muandet, K.; Balduzzi, D.; and Schölkopf, B. 2013. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, 10–18.
- Orlando, J. I.; et al. 2020. Refugee challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59: 101570.
- Pandey, P.; Raman, M.; Varambally, S.; and AP, P. 2021. Generalization on Unseen Domains via Inference-Time Label-Preserving Target Projections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12924–12933.
- Price, W. N.; and Cohen, I. G. 2019. Privacy in the age of medical big data. *Nature medicine*, 25(1): 37–43.
- Qiao, F.; Zhao, L.; and Peng, X. 2020. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12556–12565.
- Seo, S.; Suh, Y.; Kim, D.; Han, J.; and Han, B. 2020. Learning to optimize domain specific normalization for domain generalization. *ECCV*.
- 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.
- Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A.; and Hardt, M. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, 9229–9248. PMLR.
- Tong, T.; Wolz, R.; et al. 2015. Discriminative dictionary learning for abdominal multi-organ segmentation. *Medical image analysis*, 23(1): 92–104.
- Volpi, R.; Namkoong, H.; Sener, O.; Duchi, J. C.; Murino, V.; and Savarese, S. 2018. Generalizing to unseen domains via adversarial data augmentation. In *Advances in neural information processing systems*, 5334–5344.
- Vu, T. H.; and Monga, V. 2017. Fast low-rank shared dictionary learning for image classification. *IEEE Transactions on Image Processing*, 26(11): 5160–5175.
- Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2021a. Fully test-time adaptation by entropy minimization. *International Conference on Learning Representations*.
- 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.
- Wang, Z.; Luo, Y.; Qiu, R.; Huang, Z.; and Baktashmotlagh, M. 2021b. Learning to Diversify for Single Domain Generalization. *arXiv preprint arXiv:2108.11726*.
- Yue, X.; Zhang, Y.; Zhao, S.; Sangiovanni-Vincentelli, A.; Keutzer, K.; and Gong, B. 2019. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE International Conference on Computer Vision*, 2100–2110.
- 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*.
- Zhang, S.; Zhan, Y.; Dewan, M.; Huang, J.; Metaxas, D. N.; and Zhou, X. S. 2012. Towards robust and effective shape modeling: Sparse shape composition. *Medical image analysis*, 16(1): 265–277.