From Denoising Training to Test-Time Adaptation: Enhancing Domain Generalization for Medical Image Segmentation

Ruxue Wen Hangjie Yuan Dong Ni* Wenbo Xiao Yaoyao Wu
Zhejiang University
Hangzhou, Zhejiang, China
{ruxue.wen, hj.yuan, dni, xiaowenbo, yaoyaowu}@zju.edu.cn

Abstract

In medical image segmentation, domain generalization poses a significant challenge due to domain shifts caused by variations in data acquisition devices and other factors. These shifts are particularly pronounced in the most common scenario, which involves only single-source domain data due to privacy concerns. To address this, we draw inspiration from the self-supervised learning paradigm that effectively discourages overfitting to the source domain. We propose the Denoising Y-Net (DeY-Net), a novel approach incorporating an auxiliary denoising decoder into the basic U-Net architecture. The auxiliary decoder aims to perform denoising training, augmenting the domain-invariant representation that facilitates domain generalization. Furthermore, this paradigm provides the potential to utilize unlabeled data. Building upon denoising training, we propose Denoising Test Time Adaptation (DeTTA) that further: ***(i)*** *adapts the model to the target domain in a sample-wise manner, and* ***(ii)*** *adapts to the noise-corrupted input. Extensive experiments conducted on widely-adopted liver segmentation benchmarks demonstrate significant domain generalization improvements over our baseline and state-of-the-art results compared to other methods. Code is available at https://github.com/WenRuxue/DeTTA.*

1. Introduction

In the last decade, deep learning has been extensively studied for assisting medical image analysis, aiming to reduce doctors' workload. Medical image segmentation, a critical prerequisite for various clinical analyses, has received significant attention. Many deep neural networks [26], represented by U-Net [37], demonstrate remarkable performance in various medical image segmentation tasks. However, in clinical practice, medical images often display distributional discrepancies due to factors such as different equipment, diverse imaging parameters, and fluctuations in signal-to-noise ratio over time. While cross-modality datasets exhibit greater domain shift (e.g., from MRI to CT), it is important to note that scenarios involving the same modality are more prevalent in clinical practice. Therefore, we restrict our focus solely to discrepancies in cases involving the same modality. Such discrepancies challenge deep neural networks to generalize to unseen domains, leading to performance degradation.

To address this problem, domain generalization (DG) has emerged to enhance the generalization ability of deep neural networks to unseen domains. Most existing domain generalization methods [23,32,51,56] aim to achieve generalization performance in unseen domains by extracting domain-invariant features from multiple domains [27]. These methods prove ineffective when dealing with the problem of medical image segmentation due to the scarcity of available data and available data annotations [27].

In medical image segmentation, a more realistic yet challenging setting is single domain generalization (SDG), where only one single domain is available for training. For the challenging SDG problem, an intuitive solution is to increase the diversity of training data through adversarial data augmentation [34,43,52] or data generation [24,35,39,53]. However, synthesizing high-quality medical images with intricate details is challenging, and these methods often struggle to perform well in domains that differ significantly from the source domain due to the challenge in anticipating the distribution of test data [27]. In addition to data manipulation, SDG is also studied in general machine learning paradigms [45], such as dictionary learning [27], contrastive learning [9,13,18]. Furthermore, there are also several studies embarking on the exploration of leveraging self-supervised learning, such as predicting the shuffling order of patch-shuffled images [28] or rotation degrees [11], to enhance domain generalization performance. An intuitive explanation is that the self-supervised learning paradigm allows a model to learn generic features and reduces the like-

[*] Corresponding author.

lihood of overfitting to the source domain [6].

Inspired by the success of self-supervised learning for generalization, we aim to address the SDG problem in a novel way for medical image segmentation. We observe that medical images often suffer from various types of noise due to limitations in imaging technology or variations in imaging protocols, which is one of the key factors contributing to domain shift [16]. Hence, properly leveraging self-supervised denoising can disregard the noise in medical images from different domains, allowing the network to focus more on clean images. Secondly, self-supervised denoising benefits from all available raw images, thereby enhancing the feature extraction capabilities of the encoder [5]. Thirdly, the single given test data hints at its distribution, enabling us to adapt the model to each unlabeled test data at test time only. This test-time adaptation approach is compatible with solving the SDG problem [25].

With these insights, we present Denoising Y-Net (DeY-Net), a novel approach with a Y-shaped architecture to enhance domain generalization for medical image segmentation. DeY-Net consists of an encoder followed by two decoders: a decoder for pixel-wise segmentation and an auxiliary decoder for self-supervised denoising, utilizing the Noise2Void training scheme [20]. Furthermore, the self-supervised denoising branch provides the potential to utilize unlabeled data. Building upon denoising training, we propose Denoising Test Time Adaptation (DeTTA) that further: **(i)** adapts the model to the target domain in a sample-wise manner, and **(ii)** adapts to the noise-corrupted input, achieving more stable performance improvements.

Our main contributions are highlighted as follows:

- We present a novel architecture named DeY-Net, to address the SDG problem by incorporating a self-supervised denoising decoder into a basic U-Net.
- We propose Denoising Test-Time Adaptation (DeTTA), which adapts the model to the target domain and adapts to the noise-corrupted input in order to preserve more information.
- We conduct extensive experiments on a widely-adopted liver segmentation task. By training only on a single domain, our method significantly improves generalization performance over our baseline and state-of-the-art results compared to other methods.

2. Related work

2.1. Denoising for segmentation

It is well-known that there is an overlap between denoising and segmentation tasks [49]. Various works have proved that the self-supervised denoising task can enable the segmentation task, especially in the presence of extreme levels of noise and limited training data [33]. Mangal Prakash *et al*. [33] demonstrate that the self-supervised denoising prior [20] can significantly improve segmentation results, where denoising and segmentation are realized in two sequential steps. Similarly, Sicheng Wang *et al*. [47] use a network that incorporates tandem segmentation and denoising tasks. Tim-Oliver Buchholz *et al*. [5] propose using a single network to jointly predict the denoised image and the desired object segmentation. Emmanuel Asiedu Brempong *et al*. [4] propose Decoder Denoising Pretraining (DDeP) to pretrain the segmentation decoder with a well-trained denoising network. These approaches utilize self-supervised denoising to improve segmentation performance via network architecture or training schemes. Building on this inspiration, we present a novel Y-shaped architecture integrating self-supervised denoising as a secondary decoder in the network.

2.2. Self-supervised learning

Self-supervised learning (SSL) is a learning paradigm that enables learning semantic features by generating supervisory signals from a pool of unlabeled data [38]. In the medical field, several works have demonstrated that SSL can produce a pretrained model to advance supervised tasks [14], such as image classification [1, 14] and image segmentation [15, 29]. These approaches learn image representations through handcrafted pretext tasks [38] such as image rotation prediction [11], in-painting [31], Jigsaw puzzle [28], denoising auto-encoder [42] and so on. Besides, for generalization purposes, some methods utilize self-supervised learning as an auxiliary task for the main task [6]. As early as 2016, Muhammad Ghifary *et al*. [10] propose to add a reconstruction decoder that shares the encoding representation with the classification head, which can be trained with unlabeled target domain data. Inspired by this, Joris Roels *et al*. [36] propose a domain adaptation (DA) method, named Y-Net, by integrating a reconstruction decoder for medical image segmentation within the Y-shaped architecture. Y-Net is also proposed in [46]. The distinction between DA and DG lies in their utilization of target domain data during the training phase, which is exclusive to DA and not employed in DG. Kai Zhu *et al*. [55] devise a Self-Supervised Module (SSM) to improve the segmentation performance. Yu Sun *et al*. [40] integrate a self-supervised image rotation classifier head, allowing for the utilization of the unlabeled test data at test time. Inspired by these works, we further explored the effectiveness of incorporating a self-supervised branch to enhance the model's generalization performance.

2.3. Test-time adaptation

In recent years, there has been significant development in Test-Time Adaptation (TTA). TTA aims to utilize the distribution information from the test data to quickly adapt mod-

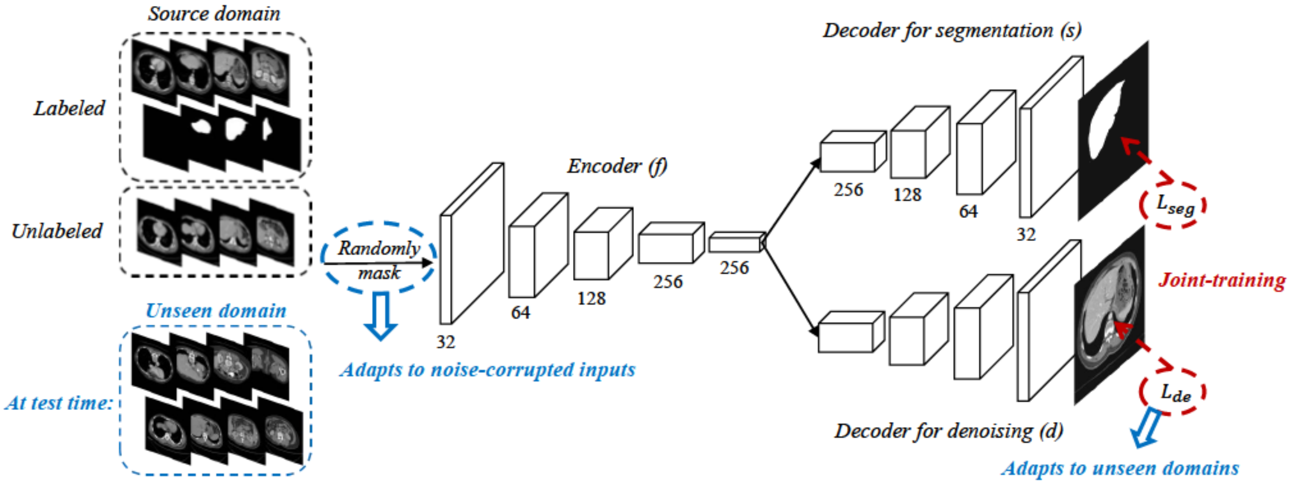

Figure 1. Overview of the proposed DeY-Net. The fundamental backbone is U-Net, and the structures of the two decoders are identical and retain the skip connections (omitted in the figure). We joint-train the segmentation and the self-supervised denoising tasks by combining their losses L_{seg} and L_{de}. When testing data on a new domain, we only adapt the encoder parameters through the self-supervised denoising branch with L_{de} to improve generalization. Besides, each test image is randomly masked with neighbor pixels several times to get a more stable average prediction.

els with a few gradient steps [7, 27, 30]. The main differences of prior work lie in how to devise the objective that can be optimized with unlabeled test data and which part of network parameters to be updated at test time [27]. For instance, TTT [40] adapts the encoder of the classification model at test time via an auxiliary branch with rotation prediction self-supervision. Later on, Tent [44] optimizes entropy minimization loss of predictions on test data to adapt the batch normalization layer. For test-time adaptation on medical image segmentation, Hu *et al.* [12] propose using new losses like Regional Nuclear-norm (RN) and Contour Regularization (CR) losses to improve generalization performance. Neerav Karani *et al.* [16] propose to generate pseudo-labels through a denoising autoencoder at test time to adapt an image normalization module. The denoising autoencoder is a separate network that needs to be trained independently from the segmentation network. Similarly, Jeya Maria Jose Valanarasu *et al.* [41] also propose to train an additional autoencoder to adapt the Adaptive Instance Norm (AdaIN) layers, which profoundly relies on the training performance of the additional autoencoder network. In contrast, our work utilizes a single network, simplifying the training process and ensuring consistent and reliable results.

3. Methodology

In this section, we first provide an overview of the SDG problem and our proposed DeY-Net. We then introduce the basic principle of the self-supervised denoising scheme Noise2Void [20] as the preliminary of DeY-Net and explain denoising training and test-time adaptation (DeTTA) in detail.

3.1. Overview

In the setting of single domain generalization (SDG) [45], we are given only one training (source) domain and we denote it as $S_{train} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n} \sim P_{XY}$, where $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$ denotes the input, $y \in \mathcal{Y} \subset \mathbb{R}$ denotes the label, and P_{XY} denotes the joint distribution of the input sample and output label. X and Y denote the corresponding random variables. The goal of SDG is to learn a robust and generalizable predictive function $h : \mathcal{X} \rightarrow \mathcal{Y}$ from the single source domain to achieve a minimum prediction error on an unseen test domain S_{test} (i.e., S_{test} cannot be accessed in training and $P_{XY}^{test} \neq P_{XY}^{train}$):

$$\min_{h} \mathbb{E}_{(x,y)\in S_{test}}[l(h(\mathbf{x}), y)], \tag{1}$$

where $\mathbb{E}$ is the expectation and $l(\cdot,\cdot)$ is the loss function.

On this basis, test-time adaptation (TTA) utilizes the unlabeled test sample $\mathbf{x}_i^{test} \in S_{test}$ presented at test time to adapt the model for generalization purpose.

To address the SDG problem, an overview of our method DeY-Net is illustrated in Fig. 1. While we build upon the U-Net architecture, our approach utilizes a Y-shaped design, where the encoder is followed by two decoders: the segmentation decoder for pixel-wise segmentation and the denoising decoder for self-supervised denoising.

3.2. Preliminary of DeY-Net

The core idea of DeY-Net revolves around utilizing self-supervised denoising to enhance generalization performance. Before delving into the details of our approach, the principles of the self-supervised denoising will be introduced. We use the Noise2Void (N2V) scheme described

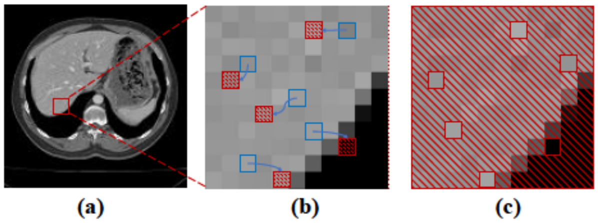

Figure 2. The training scheme of Noise2Void. **(a)** An original training image. **(b)** A magnified image patch extracted from (a). Similar operations are performed throughout the entire image. During N2V training, several random pixels (red and striped squares) are replaced by neighboring pixels (blue squares). This modified image is then used as the input image. **(c)** The target patch corresponding to (b). The loss is only calculated for the pixels masked in (b).

in [20] as our self-supervised denoiser of choice. Conveniently, N2V uses a default U-Net with a modified input and loss for denoising. We replicate the decoder of the original U-Net to serve as the second decoder for N2V.

In N2V, the noise is assumed pixel-wise independent. Thus the noise information is not carried in the neighboring pixels. Hence, denoising is feasible by predicting the pixels' values replaced by neighboring pixel values. As the training scheme shown in Fig. 2, N2V randomly selects N pixels in each training image; then these pixels are replaced with neighboring pixels. The training targets are the corresponding original pixel values.

The N2V network can be trained by minimizing the empirical risk

$$\arg\min_{f} \sum_{j} \sum_{i} L(f(\tilde{x}^{j}_{RF(i)}), x^{j}_{i}), \tag{2}$$

where $\tilde{x}^{j}_{RF(i)}$ is a patch around pixel i, extracted from training input image x^j. In this patch, the value at the position i is replaced with the value of a neighboring pixel. x^j_i is the corresponding target pixel value. The summation of the losses of all N pixels i in the entire training image yields the total loss for each image.

For L, we consider the standard MSE loss:

$$L(\tilde{s}^{j}_{i}, s^{j}_{i}) = (\tilde{s}^{j}_{i} - s^{j}_{i})^2. \tag{3}$$

3.3. Training of DeY-Net

Pretraining. To utilize the efficacy of denoising, we perform a separate pretraining step for the segmentation decoder using a well-trained denoising network [4]. Specifically, we trained a U-Net with the Noise2Void training scheme, utilizing all the source domain data without any segmentation labels. Subsequently, we copy the parameters of the trained decoder to the segmentation decoder of DeY-Net, while the encoder and the denoising decoder of DeY-Net are randomly initialized. If we simultaneously pretrain other components, it may result in overfitting. In the ablation experiments, we will demonstrate the effectiveness of the pretraining step and compare the results obtained by pretraining different components separately.

Joint-training. As shown in Fig. 1, our Y-shaped architecture allows the simultaneous execution of two tasks. We train both the segmentation and the denoising tasks by summing their respective losses. The encoder, the pretrained segmentation decoder, and the denoising decoder are denoted as f, s_0, and d, respectively.

We randomly mask pixels of the input image x^j from the training set and replace them with the neighboring pixel values, getting the actual input $\tilde{x}^j$. The original input values at the masked positions are the training targets for the denoising task.

The denoising loss is evaluated for labeled images and unlabeled ones. We use the standard Noise2Void loss, which is expressed as:

$$l_{de} = \sum_{j} \sum_{i} L(d \circ f(\tilde{x}^{j}_{i}), x^{j}_{i}). \tag{4}$$

To address the common imbalance in the number of foreground and background pixels processed in medical images, we choose the standard Dice loss [3] as the supervised segmentation loss, evaluated for labeled images only, with labels as $\mathbf{y}$. The segmentation loss can be expressed as:

$$l_{seg} = \sum_{j} Dice(s_0 \circ f(\tilde{x}^{j}), y^{j}). \tag{5}$$

In contrast to the straightforward joint-training process in TTT [40], our method incorporates an enhanced joint-training process by introducing a time-dependent weight to better combine supervised and unsupervised loss [21]. The joint-training produces a trained encoder f_0 and two trained decoders s_1 and d_0:

$$f_0, d_0, s_1 = \arg\min_{f,d,s}(l_{seg} + w(t) * l_{de}). \tag{6}$$

In our implementation, the weighted function $w(t)$ of unsupervised loss slopes upwards from 0 along the Gaussian curve for the first 200 training periods. It means that the denoising decoder, which assists the encoder in feature extraction, slowly starts to work during the training process. Initially, the model training is primarily driven by the segmentation loss, ensuring the model does not converge to a degenerate solution where meaningful segmentation is not achieved [21]. This weight adjustment allows the main segmentation task and the auxiliary denoising task to strike a balance, ensuring practical completion of the segmentation task while preserving the generalization ability offered by the denoising task.

3.4. DeTTA

Target domain adaptation. At test time, considering that the single test volume can give us a hint about its distribution, we aim to optimize the model parameters with each unlabeled test volume. Once each test volume x arrives, we optimize the denoising loss (Eq. (4)) on the denoising branch $d_0 \circ f_0$, while the segmentation decoder s_1 is frozen:

$$f_x, d_x = \arg\min_{f,g} L(d_0 \circ f_0(\tilde{x}), x). \tag{7}$$

We only perform a one-step gradient descent on each test volume, a medical volume data from a specific patient. Using the whole volume of data simultaneously is consistent with the clinical scenario where test data usually arrives per patient. To preserve the original discriminability of the model, we only adapt the parameters in the batch normalization layers for the test-time adaptation. It is motivated by the fact that modifying all the parameters of the model is unstable and inefficient when only a single test sample is available at test time [44].

After test-time adaptation to each test volume, we get an adapted model for segmentation $s_1 \circ f_x$, and we make predictions on the test volume x as $s_1 \circ f_x(\tilde{x})$. Note that we use $\tilde{x}$ instead of x as the prediction input.

The above adaptation in our method is not performed online, as we do not assume that each medical volume data comes from the same distribution. After making predictions on each test volume x, we always discard f_x and d_x, and reset the weights to d_0 and f_0 for the next test volume.

Noise-corrupted input adaptation. Indeed, it is apparent that the modified input $\tilde{x}$ leads to the loss of information of the original test image. As mentioned in Sec. 3.3, the inputs during training are modified as $\tilde{x}$ to facilitate the joint-training of the two tasks. If we directly make predictions on original x as $s_1 \circ f_x(x)$ to preserve all the original information, the reliability of the predictions may be compromised. To further preserve the information, DeY-Net further adapts to the noise-corrupted input.

Intuitively, for each original test x, we can get several inputs $\{\tilde{x}, \tilde{x}^{'}, \tilde{x}^{''}, \cdots\}$ after being randomly masked separately. Since the positions of the masked pixels are random, such a set of inputs minimizes the information loss. Then, we make predictions on all these masked inputs and merge the predictions. It is a novel test-time augmentation strategy designed for our method. The pipeline is shown in Fig. 3.

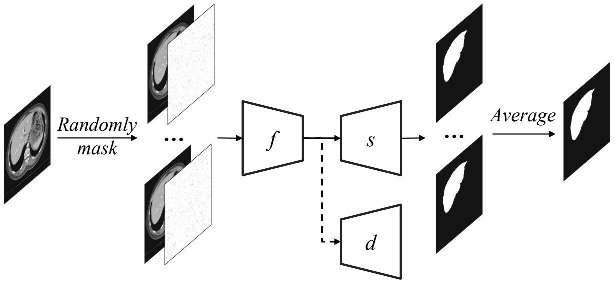

Figure 3. The pipeline of noise-corrupted input adaptation.

4. Experiment

We assess the effectiveness of the proposed DeY-Net and DeTTA on a typical medical image segmentation: liver segmentation on CT. Experimental details, comparison results with other methods, and ablation studies are detailed in the following subsections.

4.1. Experimental settings

Datasets. We use CHAOS-train [17] with 20 CT volumes as the labeled dataset and CHAOS-test [17] with 20 CT volumes as the unlabeled dataset. Both the labeled and unlabeled data are collected from healthy patients only. We further randomly split the CHAOS-train dataset into train and test sets with 16 and 4 volumes, respectively, for source domain training and in-domain testing. To assess the out-of-domain generalization, we evaluate three additional out-of-domain datasets, including LITS2017 [2] and two datasets from local clinical centers, named Normal and Ill. LITS2017 is a challenging dataset that contains 130 CT volumes from healthy patients and patients with liver tumors. Normal contains 30 CT volumes from healthy patients, while Ill contains 15 from patients with cirrhosis, both annotated by experts.

These datasets constitute at least 4 data domains, of which LITS2017 collected by multiple clinical centers, simply handled as a data domain. Fig. 4 shows the representation cases and volume number of each dataset, showing the domain differences (such as the liver's position, resolution, and direction). In addition, liver morphology varies between patients with cirrhosis, tumors, and healthy patients. Since the thickness of data slices varies greatly between clinical sites, we validated our method on these data using a 2D network as the backbone network.

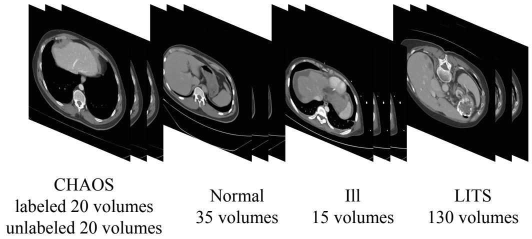

Figure 4. The representative cases and the volume number of 4 datasets.

Preprocessing. LITS2017 is a collection of liver CTs with liver and tumor segmentation labels. We only use the liver class and merge the tumor class into the liver class. The preprocessing in [22] is used for all the raw CT data. We first clip the attenuation coefficient in a range between -200 and 400, highlighting the liver portion, and subsequently normalize it by subtracting the minimum and dividing by the signal range in a slice-wise manner. To process the input, we set the masked ratio at 0.1, meaning that 10% pixels of each input are selected and replaced with the

Dataset	in-domain	out-of-domain			
	CHAOS test	Normal	Ill	LITS	average
	Dice Coefficient↑ (Dice, mean±std)				
U-Net [37]	96.32±0.20	92.71±0.90	88.75±0.51	85.33±0.29	90.78
BigAug [50]	96.34±0.05	93.08±0.18	89.21±0.80	85.28±0.31	90.98
DualNorm [54]	96.20±0.16	94.09±0.03	90.22±0.28	83.54±0.23	91.01
TTT [40]	95.62±0.43	94.57±0.31	90.54±0.40	84.45±0.27	91.61
Tent [44]	96.55±0.04	95.24±0.04	90.79±0.46	85.55±0.31	92.03
RN+CR [12]	96.59±0.08	95.26±0.04	90.76±0.47	85.58±0.33	92.05
TTST [16]	96.71±0.17	95.32±0.33	91.48±0.26	85.67±0.66	92.30
OtF [41]	96.22±0.01	94.45±0.03	88.14±0.08	81.17±0.01	90.00
ReY-Net (*w/o* ReTTA)	96.57±0.06	95.72±0.04	90.44±0.51	82.35±2.76	91.27
ReY-Net (*w/* ReTTA)	96.23±0.95	95.51±0.16	90.93±0.26	86.34±0.36	92.26
DeY-Net (*w/o* DeTTA)	96.63±0.03	**95.74±0.04**	91.50±0.18	85.19±0.04	92.27
DeY-Net (*w/* DeTTA)	**96.71±0.12**	95.66±0.08	**91.60±0.14**	**87.14±0.09**	**92.77**

Table 1. Quantitative comparison of domain generalization results.

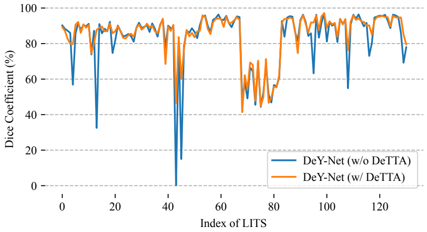

Figure 5. DeTTA improvement of LITS.

Dataset	LITS	Tumor
DeY-Net (*w/o* DeTTA)	85.19	83.92
DeY-Net (*w/* DeTTA)	87.14	88.51
△	**1.95**	**4.59**

Table 2. Quantitative results of DeTTA improvement. Tumor datasets consists of images containing tumors in the LITS dataset.

neighboring pixel values following N2V. We do not do any cropping, resampling, or alignment but only the slice-wise preprocessing described above.

Metrics. For evaluation, we use a commonly used metric, the Dice coefficient [%] [8], to quantitatively evaluate the segmentation results. To ensure the stability and reliability of our experiments, we conduct each experiment three times utilizing different random seeds. The results are presented using the average value with the standard deviation.

4.2. Implementation details

In our experiment, the test-time adaptation updates just a step with a learning rate of 1e-6. For the test-time augmentation, we augment twice and average the two outputs. We use the U-Net as the backbone for all experiments. No data augmentation technology is utilized in our methods and other comparison methods. In our method, the U-Net is also used for the pretrained denoising model, which is trained with all the CHAOS data of 200 epochs. Both the pretrained denoising U-Net and our proposed model DeY-Net are trained using Adam optimizers [19] with the momentum of 0.9 and 0.99, and the learning rate is initialized to 1e-4. We train the DeY-Net for 200 epochs on the CHAOS dataset only until convergence. The batch size of the baseline U-Net and the DeY-Net are consistently set to 8, which means four labeled data and four unlabeled data in each batch of DeY-Net. During training, the weighted function $w(t)$ of unsupervised loss slopes upwards from 0 along the Gaussian curve for the first 200 training periods. The maximum value of $w(t)$ is $\alpha(n_l/(n_l + n_{un}))$, where n_l and n_{un} are numbers of labeled and unlabeled data, respectively. After conducting exploratory experiments, we set the α to 30 in our experiments. The framework is implemented via Pytorch using an NVIDIA P100 GPU.

4.3. Results

Comparison methods. We compare our methods with current state-of-the-art methods to solve the SDG problem, including: **BigAug (2020)** [50], a DG method in medical image segmentation with extensive data transformations to promote general representation learning. **DualNorm (2022)** [54], an SDG method in medical image segmentation via style augmentation and dual normalization. **TTT (2020)** [40], a test-time training method in image classifier with an auxiliary self-supervised task of rotation prediction. **Tent (2021)** [44], a fully test-time adaptation method in the image classifier field by minimizing the entropy of predictions. **TTST (2021)** [16], a test-time adaptation in medical image segmentation with an additional denoising autoencoder. **RN+CR (2021)** [12], a fully test-time adaptation method in medical image segmentation by minimizing new losses, Regional Nuclear-norm (RN) and Contour Regularization (CR). **OtF (2022)** [41], an on-the-fly test-time adaptation method with an additional Domain Prior Generator.

The **Baseline** setting in domain generalization denotes learning a model (U-Net [37]) on the source domain without using any generalization technique and directly making predictions on the target domains.

Quantitative comparison results. Tab. 1 shows the results on the liver segmentation. For the comparison methods using TTA, we also perform a one-step gradient descent for each testing volume. First, we compare our method DeY-Net (*w/* DeTTA) against the baseline U-Net. While the U-Net achieves satisfactory in-domain performance but struggles with out-of-domain data, our proposed method further enhances in-domain performance and significantly improves performance on unseen domains, with an improvement of 1.99% on average.

Most of the methods can improve the generalization performance over baseline. Among all these methods, our DeY-Net (*w/* DeTTA) performs better than other methods. One noticeable observation is the significant improvement achieved by our method compared to OtF. Due to removing the back-propagation, OtF heavily relies on the generalization performance of the additional Domain Prior Generator, highlighting our joint-training approach's advantage.

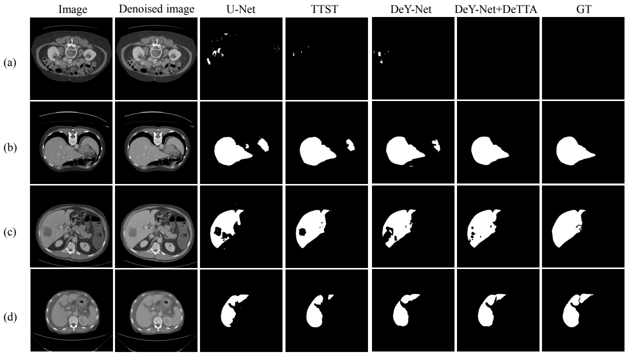

Figure 6. Visualization comparison of segmentation results with baseline methods, TTST and the proposed method DeY-Net. The samples represent the following: **(a)** A liver-free image; **(b)** A healthy liver image; **(c)** An image with liver tumors; **(d)** An image with liver cirrhosis.

Compared with BigAug and DualNorm, the DG methods that rely on data augmentation, and TTT and Tent, the TTA methods that are not designed for the image segmentation task, our method has a more significant improvement, especially on the most challenging LITS dataset. Even though RN+CR, a method similar to Tent, is designed for image segmentation, only a slight improvement over Tent is achieved in the liver segmentation experiments.

Notably, our method demonstrates the slightest improvement over TTST. This can be attributed to using an additional denoising autoencoder trained outside the segmentation network in TTST to refine the segmentation results. The sequential structure connection between the two networks consumes computational resources and adds complexity to the overall process. Instead of the sequential structure in TTST, we employ an alternative and more convenient Y-shaped architecture for leveraging self-supervised denoising, which has resulted in superior performance.

Within these out-of-domain datasets, the results of LITS are more representative of generalization ability, which contains the largest and most diverse data. These comparison methods fail in LITS dataset on average. Meanwhile, our method still improves performance on the LITS dataset, demonstrating its effectiveness even in challenging unseen domains.

Visualization comparison results. Fig. 6 further shows the segmentation results with four samples from unseen domains for the liver segmentation task. It is observed that our method demonstrates superior segmentation accuracy in unseen domains, whereas other methods may fail at times. Our method accurately segments the regions as non-liver, which are erroneously segmented as liver by other methods. Furthermore, for the regions containing the liver, our approach captures finer details and accurately delineates the liver's boundaries. In the case of CT images from patients with liver cirrhosis and liver tumors, all methods struggle to completely overcome domain shifts caused by changes in liver texture and structure. Nonetheless, our method consistently achieves more reliable segmentation results, benefiting from its effective image information utilization. This improvement illustrates the role of the denoising branch in capturing intricate details of the images.

Analysis of DeTTA improvement. Overall, DeTTA increases the average Dice by 0.50% over DeY-Net(*w/o* DeTTA), with a notable increase of 1.95% in the LITS dataset. This observation reflects that DeTTA gains additional capacity and enables the model to utilize the test data information to improve model generalizability adaptively.

However, TTA methods are not always work due to varying degrees of domain shifts [48]. Analyzing scenarios in which DeTTA encounters limitations is crucial. Our detailed analysis, conducted primarily on the challenging LITS dataset (Fig. 5 and Tab. 2), reveals that for certain test data with large domain shift and poor segmentation results, DeTTA can extract information from the test data and adapt the model according to the test domain to significantly improve the segmentation results. This is quantitatively demonstrated by the improvement of Tumor dataset(4.59%), a subset of the LITS dataset(1.95%) with more pronounced domain shifts. Conversely, when the initial segmentation results are already satisfactory, indicating a slight shift between the test data and the source domain distribution, employing DeTTA may yield slight improvement or even lead to a decline in performance. We attribute the observed decline to the potential over-optimization of the trained model to the tested sampling, leading to performance degradation even with a single step of gradient update. We anticipate that this can be decently resolved by an adaptive DeTTA strategy, which is beyond the scope of this work and left for future exploration.

4.4. Ablation study

We conduct ablation studies about several key points in our model: (1) the contribution of DeTTA in our method; (2) the effect of the weighted function $w(t)$; (3) the effect of the optimized layers; (4) the effect of denoising pretraining; (5) the effect of data augmentation times.

Effect of denoising task. Tab. 1 shows the effect of different self-supervised tasks on the final results. TTT employs image rotation prediction as the self-supervised task. ReY-Net extends from our DeY-Net (with consistent network architecture, training techniques, etc.). and the self-supervised task strategically opts for the reconstruction most pertinent to denoising. ReTTA, accordingly, employs the reconstruction loss as its optimization objective. The outcomes demonstrate that self-supervised denoising is more conducive than alternative approaches for the main task of medical image segmentation.

Effect of $w(t)$. As shown in Fig. 7, $w(t)$ is effective for performance improvement of DeY-Net, which means that

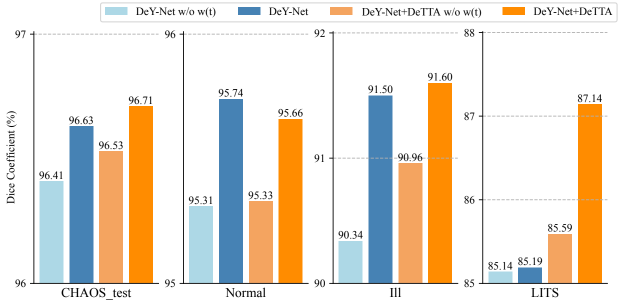

Figure 7. Ablation analysis of time-dependent weight. We analyze the effectiveness of the time-dependent weight on models with and without DeTTA.

Optimized layers	Normal	Ill	LITS
BN	**0.9566**	**0.9160**	**0.8714**
All	0.9566	0.9159	0.8696

Table 3. Ablation analysis of the optimized layers.

gradually increasing the supervision of unsupervised loss during the training process is beneficial. Notably, no matter how well the DeY-Net is trained, after the DeTTA, the overall performance will still be improved.

Effect of the optimized layers. We explore which layers need to be adapted when a model is transferred to the unseen domain. In the DeTTA, only the encoder can be optimized for segmentation, and the segmentation decoder is frozen. As shown in Tab. 3, adapting all encoder parameters is worse than adapting BN layers. It is intuitive that all parameters are adapted easily to damage the original performance of the network due to the network being over-parameterized.

Effect of denoising pretraining. As shown in Tab. 4, we conduct comprehensive experiments regarding the initialization methods for DeY-Net. The results represent the average Dice score across the four datasets. The (3) initialization method performs best when only the segmentation decoder is pretrained using a denoising U-Net. This particular pretraining approach outperforms other pretraining methods since joint-training is one key aspect of our approach. Exploiting the overlap between segmentation and denoising tasks, we pretrain the segmentation decoder to ease the segmentation task's complexity. Meanwhile, we refrain from pretraining the denoising branch to prevent premature convergence, enabling more effective joint-training of the two tasks. Compared to the random initialization (1), pretraining the segmentation decoder allows for better utilization of the denoising parameters trained on labeled and unlabeled data.

idx	f	s	d	*w/o* DeTTA	*w/* DeTTA
(1)	✗	✗	✗	0.9169	0.9246
(2)	✓	✗	✗	0.9154	0.9211
(3)	✗	✓	✗	**0.9227**	**0.9277**
(4)	✗	✗	✓	0.9206	0.9215
(5)	✓	✓	✗	0.9177	0.9208
(6)	✓	✗	✓	0.9154	0.9170
(7)	✗	✓	✓	0.9183	0.9233
(8)	✓	✓	✓	0.9084	0.9185

Table 4. Ablation analysis of the pretraining. The encoder, the segmentation decoder, and the denoising decoder are denoted as f, s, and d, respectively. ✓ represents pretraining, while ✗ represents random initialization. Our pretraining method is (3), while the baseline is (1).

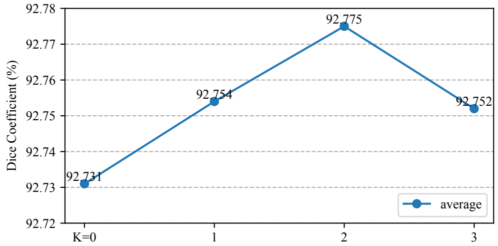

Figure 8. Ablation analysis of data augmentation times K.

Effect of data augmentation times. The choice of data augmentation times K is important in our method, affecting both the final performance and the testing efficiency. To investigate the suitable choice of K, we repeat the experiment of the DeY-Net by varying $N \in (0, 1, 2, 3)$. $K = 0$ represents the original input without being randomly masked. $K = 1$ represents the scenario without adaptation to the noise-corrupted input (data augmentation). As shown in Fig. 8, the models with data augmentation times (K=2) perform better than those with smaller or larger times on the segmentation task. We finally adopt K=2 in our method.

5. Conclusion

This paper proposes Denoising Y-Net (DeY-Net) to address the challenging SDG problem in medical image segmentation. The idea is to incorporate an auxiliary denoising decoder into a basic U-Net architecture, that naturally allows for semi-supervised training and shows strong generalization capabilities. Further, we propose Denoising Test Time Adaptation (DeTTA) to adapt the model to the target domain and adapt to the noise-corrupted input, which can further promote the model generalization at any unseen data distributions. We validate our method in the liver segmentation task. Quantitatively, we significantly outperform our baseline and other methods in- as well as out-of-domain. DeTTA may over-optimize the test data in some scenarios, leading to performance degradation even with a single step of gradient update. Also, the network architecture of our method is simple and can cope well with the domain generalization problem in the same modality case. Therefore, the future direction of research lies in developing an adaptive DeTTA strategy and cross-modality generalization.

Acknowledgement

This research was supported by the Horizontal Project of Zhejiang University (K20220130).

References

[1] Shekoofeh Azizi. Self-supervised learning advances medical image classification, 10 2021. 2

[2] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023. 5

[3] Gilles Blanchard, Aniket Anand Deshmukh, Ürun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, 22(1):46–100, 2021. 4

[4] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Denoising pretraining for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4175–4186, 2022. 2, 4

[5] Tim-Oliver Buchholz, Mangal Prakash, Deborah Schmidt, Alexander Krull, and Florian Jug. Denoiseg: joint denoising and segmentation. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, pages 324–337. Springer, 2021. 2

[6] Fabio M. Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[7] Cheng Chen, Quande Liu, Yueming Jin, Qi Dou, and Pheng-Ann Heng. Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 225–235. Springer, 2021. 3

[8] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 6

[9] Thomas Duboudin, Emmanuel Dellandréa, Corentin Abgrall, Gilles Hénaff, and Liming Chen. Encouraging intra-class diversity through a reverse contrastive loss for single-source domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 51–60, 2021. 1

[10] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 597–613. Springer, 2016. 2

[11] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 1, 2

[12] Minhao Hu, Tao Song, Yujun Gu, Xiangde Luo, Jieneng Chen, Yinan Chen, Ya Zhang, and Shaoting Zhang. Fully test-time adaptation for image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 251–260. Springer, 2021. 3, 6

[13] Shishuai Hu, Zehui Liao, and Yong Xia. Devil is in channels: Contrastive single domain generalization for medical image segmentation. *arXiv preprint arXiv:2306.05254*, 2023. 1

[14] Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P Lungren, Serena Yeung, and Akshay S Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digital Medicine*, 6(1):74, 2023. 2

[15] András Kalapos and Bálint Gyires-Tóth. Self-supervised pretraining for 2d medical image segmentation. In *European Conference on Computer Vision*, pages 472–484. Springer, 2022. 2

[16] Neerav Karani, Ertunc Erdil, Krishna Chaitanya, and Ender Konukoglu. Test-time adaptable neural networks for robust medical image segmentation. *Medical Image Analysis*, 68:101907, 2021. 2, 3, 6

[17] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021. 5

[18] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9619–9628, 2021. 1

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[20] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2129–2137, 2019. 2, 3, 4

[21] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 4

[22] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8300–8311, 2021. 5

[23] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 1

[24] Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xiaoyue Mi, Zhengze Yu, Xiaoya Li, and Boyang Xia. Progressive domain expansion network for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 224–233, 2021. 1

[25] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *arXiv preprint arXiv:2303.15361*, 2023. 2

[26] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 1

[27] Quande Liu, Cheng Chen, Qi Dou, and Pheng-Ann Heng. Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1756–1764, 2022. 1, 3

[28] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 1, 2

[29] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervised learning for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(7):1837–1848, 2022. 2

[30] Prashant Pandey, Mrigank Raman, Sumanth Varambally, and Prathosh Ap. Generalization on unseen domains via inference-time label-preserving target projections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12924–12933, 2021. 3

[31] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[32] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 1

[33] Mangal Prakash, Tim-Oliver Buchholz, Manan Lalit, Pavel Tomancak, Florian Jug, and Alexander Krull. Leveraging self-supervised denoising for image segmentation. In *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*, pages 428–432. IEEE, 2020. 2

[34] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020. 1

[35] Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan. Multi-component image translation for deep domain generalization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 579–588. IEEE, 2019. 1

[36] Joris Roels, Julian Hennies, Yvan Saeys, Wilfried Philips, and Anna Kreshuk. Domain adaptive segmentation in volume electron microscopy imaging. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1519–1522. IEEE, 2019. 2

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1, 6

[38] Saeed Shurrab and Rehab Duwairi. Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Computer Science*, 8:e1045, 2022. 2

[39] Nathan Somavarapu, Chih-Yao Ma, and Zsolt Kira. Frustratingly simple domain generalization via image stylization. *arXiv preprint arXiv:2006.11207*, 2020. 1

[40] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020. 2, 3, 4, 6

[41] Jeya Maria Jose Valanarasu, Pengfei Guo, Vibashan VS, and Vishal M Patel. On-the-fly test-time adaptation for medical image segmentation. *arXiv preprint arXiv:2203.05574*, 2022. 3, 6

[42] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 2

[43] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018. 1

[44] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 3, 5, 6

[45] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 1, 3

[46] Kaiqiang Wang, Jiazhen Dou, Qian Kemao, Jianglei Di, and Jianlin Zhao. Y-net: a one-to-two deep learning framework for digital holographic reconstruction. *Optics Letters*, 44(19):4765–4768, 2019. 2

[47] Sicheng Wang, Bihan Wen, Junru Wu, Dacheng Tao, and Zhangyang Wang. Segmentation-aware image denoising without knowing true segmentation. *arXiv preprint arXiv:1905.08965*, 2019. 2

[48] Hongzheng Yang, Cheng Chen, Meirui Jiang, Quande Liu, Jianfeng Cao, Pheng Ann Heng, and Qi Dou. Dltta: Dynamic learning rate for test-time adaptation on cross-domain medical images. *IEEE Transactions on Medical Imaging*, 41(12):3575–3586, 2022. 7

[49] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 2

[50] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J. Wood, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu.

Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Transactions on Medical Imaging*, 39(7):2531–2540, 2020. 6

[51] Fan Zhou, Zhuqing Jiang, Changjian Shui, Boyu Wang, and Brahim Chaib-draa. Domain generalization via optimal transport with metric similarity learning. *arXiv preprint arXiv:2007.10573*, 2020. 1

[52] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13025–13032, 2020. 1

[53] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 561–578. Springer, 2020. 1

[54] Ziqi Zhou, Lei Qi, Xin Yang, Dong Ni, and Yinghuan Shi. Generalizable cross-modality medical image segmentation via style augmentation and dual normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20856–20865, June 2022. 6

[55] Kai Zhu, Wei Zhai, Zheng-Jun Zha, and Yang Cao. Self-supervised tuning for few-shot segmentation. *arXiv preprint arXiv:2004.05538*, 2020. 2

[56] Ronghang Zhu and Sheng Li. Self-supervised universal domain adaptation with adaptive memory separation. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1547–1552. IEEE, 2021. 1