Curriculum-Based Augmented Fourier Domain Adaptation for Robust Medical Image Segmentation

An Wang, Mobarakol Islam, Mengya Xu, and Hongliang Ren, *Senior Member, IEEE*

***Abstract*— Accurate and robust medical image segmentation is fundamental and crucial for enhancing the autonomy of computer-aided diagnosis and intervention systems. Medical data collection normally involves different scanners, protocols, and populations, making domain adaptation (DA) a highly demanding research field to alleviate model degradation in the deployment site. To preserve the model performance across multiple testing domains, this work proposes the Curriculum-based Augmented Fourier Domain Adaptation (Curri-AFDA) for robust medical image segmentation. In particular, our curriculum learning strategy is based on the causal relationship of a model under different levels of data shift in the deployment phase, where the higher the shift is, the harder to recognize the variance. Considering this, we progressively introduce more amplitude information from the target domain to the source domain in the frequency space during the curriculum-style training to smoothly schedule the semantic knowledge transfer in an easier-to-harder manner. Besides, we incorporate the training-time chained augmentation mixing to help expand the data distributions while preserving the domain-invariant semantics, which is beneficial for the acquired model to be more robust and generalize better to unseen domains. Extensive experiments on two segmentation tasks of Retina and Nuclei collected from multiple sites and scanners suggest that our proposed method yields superior adaptation and generalization performance. Meanwhile, our approach proves to be more robust under various corruption types and increasing severity levels. In addition, we show our method is also beneficial in the domain-adaptive classification task with skin lesion datasets. The code is available at https://github.com/lofrienger/Curri-AFDA.**

Manuscript received 17 April 2023; accepted 19 June 2023. This article was recommended for publication by Associate Editor H. K. Lee and Editor B. Vogel-Heuser upon evaluation of the reviewers' comments. This work was supported in part by the Hong Kong Research Grants Council (RGC) Collaborative Research Fund under Grant CRF C4063-18G, in part by the Shun Hing Institute of Advanced Engineering (SHIAE) at The Chinese University of Hong Kong (CUHK) under Project BME-p1-21, and in part by the General Research Fund under Grant GRF 14216022. *(An Wang, Mobarakol Islam, and Mengya Xu contributed equally to this work.) (Corresponding author: Hongliang Ren.)*

An Wang and Hongliang Ren are with the Department of Electronic Engineering and the Shun Hing Institute of Advanced Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong (e-mail: wa09@link.cuhk.edu.hk; hren@cuhk.edu.hk).

Mobarakol Islam is with the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), Department of Medical Physics and Biomedical Engineering, University College London, WC1E 6BT London, U.K. (e-mail: mobarakol.islam@ucl.ac.uk).

Mengya Xu is with the Department of Biomedical Engineering, National University of Singapore, Singapore 119077, and also with the NUSRI, Suzhou 215000, China (e-mail: mengya@u.nus.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TASE.2023.3295600.

Digital Object Identifier 10.1109/TASE.2023.3295600

***Note to Practitioners*—Medical image segmentation is key to improving computer-assisted diagnosis and intervention autonomy. However, due to domain gaps between different medical sites, deep learning-based segmentation models frequently encounter performance degradation when deployed in a novel domain. Moreover, model robustness is also highly expected to mitigate the effects of data corruption. Considering all these demanding yet practical needs to automate medical applications and benefit healthcare, we propose the Curriculum-based Fourier Domain Adaptation (Curri-AFDA) for medical image segmentation. Extensive experiments on two segmentation tasks with cross-domain datasets show the consistent superiority of our method regarding adaptation and generalization on multiple testing domains and robustness against synthetic corrupted data. Besides, our approach is independent of image modalities because its efficacy does not rely on modality-specific characteristics. In addition, we demonstrate the benefit of our method for image classification besides segmentation in the ablation study. Therefore, our method can potentially be applied in many medical applications and yield improved performance. Future works may be extended by exploring the integration of curriculum learning regime with Fourier domain amplitude fusion in the testing time rather than in the training time like this work and most other existing domain adaptation works.**

***Index Terms*— Curriculum learning, Fourier transform, augmentation mixing, robustness, domain adaptive medical image segmentation.**

I. Introduction

ALTHOUGH deep learning is showing impressive performance in medical applications to boost the autonomy of computer-aided diagnosis and intervention, recent studies observe significant degradation in the deployed target dataset [1], [2], [3]. This is due to domain shifts such as population shift, covariate shift, and acquisition shift [4], [5], [6] in the deployment domain. In particular, the problem is usually unavoidable in the medical imaging field because medical data and annotations are usually limited and derived from multiple working sites with different scanners, protocols, and populations. This problem also leads to overfitting, under-specification, poor generalization and weak robustness of the model.

Many works have focused on domain adaptation (DA) and domain generalization (DG) to tackle data shifts in the target domain. Most of these works utilize supervised, semi-supervised, and unsupervised techniques with the strategies of transfer learning [7], [8], fine-tuning [9], [10], adversarial

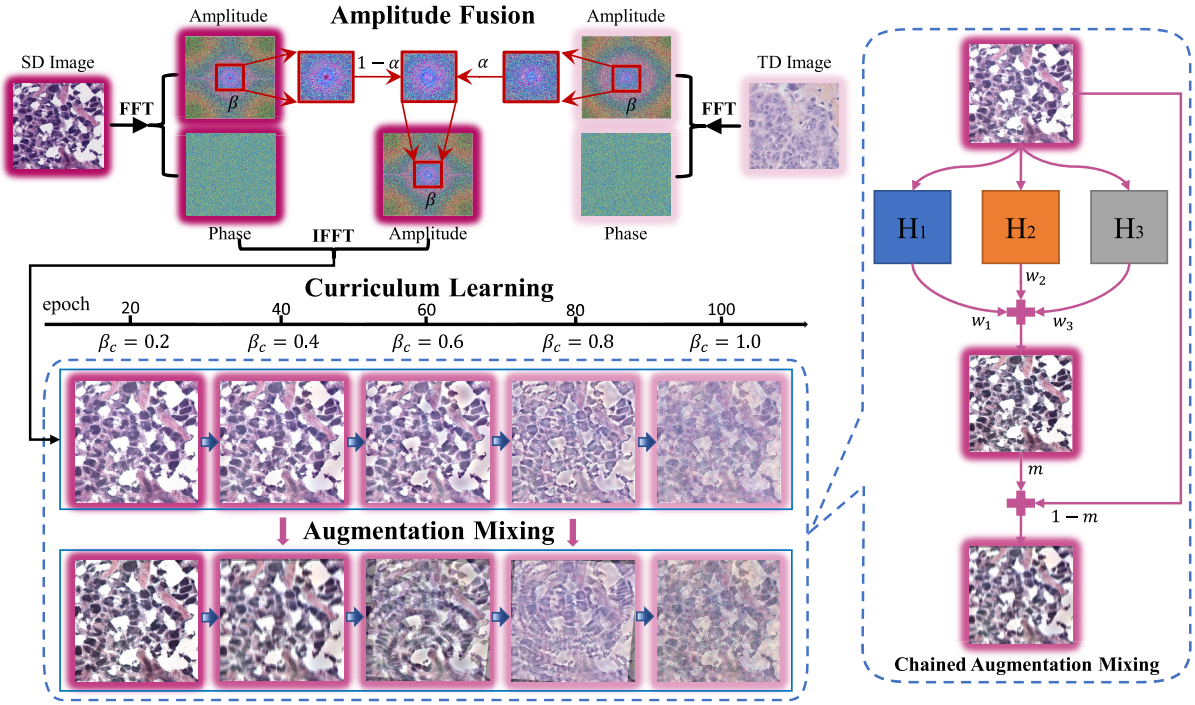


Fig. 1. **Overview of the proposed Curriculum-based Augmented Fourier Domain Adaptation (Curri-AFDA).** In the **Amplitude Fusion** (AF) module, the amplitude scaling coefficient β adjusts the central region area of the amplitude spectrum to be mixed between the source domain (SD) and the target domain (TD), and the weighting coefficient α controls the mixing strength. FFT and IFFT stand for the Fast Fourier Transform and the Inverse Fast Fourier Transform. Then the composited images are adopted in the **Curriculum Learning** (CL) process to train the domain-adaptive model. Amplitude fusion of images gradually gets enhanced when β_c linearly grows with epochs, making the source domain data appear more similar to the target domain data. During training, the **Chained Augmentation Mixing** (CAM) module helps create more variations of the training samples by mixing the outputs of up to three augmentation chains (ACs) and then with the original input image. H_i and w_i represent the sequential augmentations and the mixing weight of the i^{th} chain, respectively. m denotes the mixing weight with the original input.

training [11], [12], [13], and data augmentation [14], [15]. Depending on the availability of the target domain data during training, there are typically two types of domain adaptation. Testing-time DA, like domain generalization, tries to handle unseen domain shifts from training. Training-time DA, where target domain data is available with limited annotations (weakly-supervised DA) or no annotations (unsupervised DA), mainly emphasizes transferring target domain information to the source domain during training. The transferring methods can also be categorized as feature-level transferring [16], [17], image-level transferring [18], [19], and label-level transferring [20]. Recently, Fourier Transform has been used to transfer domain-specific information from target images to source images by performing amplitude fusion in the frequency domain [2], [21], [22], [23], [24]. These studies show the effectiveness of the Fourier technique with the advantage of simplicity and model-agnostic characteristics. However, besides adaptation and generalization, the above works seldom explore robustness under naturally-induced data alterations and corruption, which is also crucial in the model deployment phase.

Recently, curriculum learning [25], a training scheme that aims to let the model learn from easier to more complex samples or tasks, has been captivating increasing attention in the field of computer vision. One of the key benefits of curriculum learning is that it can improve a model's generalization performance. The efficacy of replacing conventional training with curriculum learning has been demonstrated in many application fields, such as semantic segmentation [16], [26], object detection [27], [28], neural machine translation [29], image captioning [30], [31], and robotic learning [32]. The efficacy of a curriculum-based model mainly depends on the proper design of the difficulty measurement process for training samples or tasks. Specifically for curriculum-based Domain Adaptation, different approaches such as domain discriminator [33], density-based clustering [34], superpixel label transfer [16], and domain similarity grouping [35] have been proposed to quantify difficulty in a weakly supervised or unsupervised manner.

The generalization and robustness abilities of the deep learning model are frequently observed to be improved by augmenting training data. Multiple augmentation techniques have been developed to boost model performance in a cutting-based or mixing-based manner, e.g., CutOut [36], MixUp [37], CutMix [38], and AugMix [39]. To assess the robustness of the deep learning model, benchmark datasets for two types of robustness (corruption and perturbation) are created [40]. The enhanced robustness is demonstrated and proven with an altered test dataset that includes corrupted and perturbed images [41], [42].

In this work, we design a Curriculum-based Augmented Fourier Domain Adaptation (Curri-AFDA) method to tackle domain shift by transferring target domain information to the source domain in a curriculum manner and extensively augmenting the data by training-time chained augmentation mixing, as shown in Fig. 1. To build the curriculum strategy,

we consider modeling the difficulty of domain adaptation as recognizing target domains with different levels of distribution shift. Specifically, we utilize Fourier Transform to extract and fuse the source and target domain information over the training period in an easier-to-harder curriculum order by progressively increasing the amplitude transferred from the target to the source domain in the frequency space. Then the reconstituted training samples are passed through chains of various augmentation operations in random orders to further improve data diversity. We validate the proposed approach on two medical image segmentation datasets of Retina and Nuclei segmentation collected from multiple domains with obvious domain shifts. We also evaluate the robustness of our method by applying 15 different corruption and perturbation techniques with five increasing severity levels on the test dataset. Extensive cross-domain validation and robustness results suggest that our approach not only improves the performance of mitigating domain variance but is also highly robust against heavy data corruptions.

Our main contributions and findings can be summarized as follows:

- Demonstrate the progressively incremental amplitude fusion in the Fourier space as an effective curriculum-based approach to alleviate domain discrepancy.
- Incorporate the training-time chained augmentation mixing to further boost the training data diversity and establish the Curriculum-based Augmented Fourier Domain Adaptation (Curri-AFDA).
- Conduct extensive experiments on multiple Retina segmentation and Nuclei segmentation datasets and various types and levels of corrupted datasets to show the superiority of our method with respect to adaptation, generalization, and robustness.
- Explore the efficacy of Curri-AFDA for the image classification task besides segmentation with the skin lesion datasets and show the potential of our method for broader medical applications.

II. Related Works

A. Fourier Transform for Domain Adaptation

Due to its simplicity, effectiveness, and model-agnostic characteristic, Fourier Transform is one of the recent tools in Domain Adaptation. In the Fourier-based frequency space, the low-frequency amplitude components, i.e., the central region of the amplitude spectrum, carry more domain-specific information. Fourier Domain Adaptation (FDA) [21] applies Fourier Transform and its inverse to spatial images and fuses the amplitude spectrum in the low-frequency region of the source domain and target domain samples to tackle domain shift. Similarly, amplitude fusion is performed by preserving the phase information for unsupervised domain adaptation [43]. Basically, the Fourier-based domain adaptation method tries to mitigate the domain gap by image-to-image translation (I2I) or style transfer - one of the major strategies for domain adaptation. After style transfer, the source domain data is expected to share a similar style as the target domain. By amplitude mixing in the frequency space, the training data appears to be in an intermediate style between the SD and TD, depending on the fusion strength. At the same time, the core domain-invariant semantics information remains unchanged in the generated image. Instead of swapping low-frequency amplitude components, Fourier augmented co-teacher (FACT) [22] and AmpMix [44] proposes to mix the whole amplitude spectrum with the MixUp [37] technique and achieves better generalization ability. By assigning pixel-wise significance with Gaussian distribution and introducing pixel-wise disturbance in the amplitude spectrum, HCDG [23] proposes to highlight the core information in the center area of the image than the marginal area. Moreover, in the federated learning scenario, Federated Domain Generalization (FedDG) [2] constructs a continuous frequency space, where low-frequency amplitude components from multiple remote domains/sites are extracted, stored, and used for training.

Compared with GAN-based domain adaptation methods, Fourier-based methods avoid additional efforts of complicated adversarial training to accomplish the domain alignment. Besides, in the case of limited training data, GAN-based methods may fail to work since they are known to be heavily data-hungry. Whereas, Fourier-based approaches are less affected and thus have significant advantages in resolving domain shift problems with insufficient data, which is meaningful for practical medical applications.

The methods mentioned above all have a fixed amplitude fusion process. For example, the portion of amplitude components to be transferred from the target domain image to the source domain image remains the same throughout the entire training. On the contrary, our proposed method introduces more target domain information to the source domain by progressively increasing the number of mixed amplitude components following the training scheduling functions. In this way, we implement a curriculum-style dynamic training scheme for the frequency domain adaptation and generate more variations of the training data in a "easier to harder" order.

B. Curriculum-Based Domain Adaptation

The curriculum-style domain adaptation approach has attracted the interest of the research community due to its excellent generalization ability. The core idea of this strategy is to learn "from easier to harder" either from the perspective of tasks or samples. At the task level, the work [45] designs the curriculum in a coarse-to-fine manner by decomposing challenging tasks into sequences of easier intermediate goals that are used to pre-train a model before tackling the target task. The efficacy of a curriculum scheme mostly depends on the appropriate difficulty measurement process. Various techniques are utilized for this purpose in domain adaptation tasks. For example, a domain discriminator to measure easier domain for multi-source domain adaptation [33], a density-based clustering algorithm to sort the samples from the target domain based on distance [34], and semantically easier class region can be considered as the easier label to train in curriculum strategy [16].

Unlike previous works, we apply curriculum-based domain adaptation by gradually introducing domain-variant information from the target domain (TD) to the source domain (SD) to mitigate the domain shift. Specifically, we take advantage of the property that the amplitude components of the frequency-domain image contain essential and specific low-level statistics. Then we design a progressive style alignment method between the source domain and the target domain by amplitude fusion of images. In such a manner, the training samples will carry more target domain information and appear more similar to the target domain images in the later training phase. By building up understanding slowly and systematically through our carefully designed curriculum, the model is able to learn more robust, generalized representations of the data. This can lead to improved performance on new, unseen data, as the model has learned to recognize more complex patterns and generalize them to new situations.

C. Data Augmentation by Mixing Images

To overcome the problem of overfitting, poor robustness, and weak generalization of deep learning models, various approaches have been proposed. Among them, data augmentation techniques, which create novel variations of the existing training images, have gained continuous attention over the years. Apart from traditional techniques like color mutation and geometric transformation, data augmentation can also be done by simply removing part of the original image [36] or further replacing it with a certain noise [38]. Except for cutting, another line of research also apply image mixing to generate new images. A pioneer mixing method is MixUp [37], followed by many other works in this area [39], [46], [47]. Among them, AugMix [39] is different in that it mixes more than two images from up to three augmentation chains. In each augmentation chain, several base augmentation operations (e.g., translation, rotation, auto-contrast) are arbitrarily applied to the original image. Then the augmented images from all chains are linearly mixed with the original image to form an overall training sample. The use of mixing augmentation can enhance the diversity of training data, which is crucial for improving robustness against unexpected shifts and corruptions in data.

Considering this, we also design the chained augmentation mixing strategy in our curriculum-based training process to enhance the generalization and robustness performance. Compared with sequentially conducting separate augmentations in a normal training scheme, our one-step chained augmentation mixing is more efficient in improving the training data diversity, only with minimal cost of matrix-weighted addition and no other computational complexity in neither the training nor test stage.

III. Methodology

In this work, we design a curriculum-based cross-domain information fusion strategy in the Fourier space and incorporate the training-time chained augmentation mixing module to improve the model performance concerning adaptation, generalization, and robustness against natural and synthetic data shifts. As shown in Fig. 1, our method mainly consists of three components: Amplitude Fusion, Curriculum Learning, and Chained Augmentation Mixing.

A. Amplitude Fusion in the Fourier Space

For a spatial-domain digital image x, we can extract the amplitude components $A(x)$ and the phase components $P(x)$ in the frequency domain with the Fourier Transform of x, i.e., $F(x)$. As the amplitude components of the Fourier Transform carry the most domain-specific information [43], [48], [49], [50], for Domain Adaptation (DA), most frequency-domain image processing techniques manipulate only the amplitude spectrum while preserving the phase spectrum as it is critical for maintaining the overall visual look of an image [48]. The pioneering work FDA [21] attempts to tackle the domain shift problem by mutating the center region of $A(x)$ from the source domain (SD) with that from the target domain (TD) in the frequency space. If A_S and A_T are denoted as the amplitude components of two random images from SD and TD, the reconstituted amplitude components in the frequency space A_S^F at the point (u, v) can be formulated as-

$$A_S^F(u,v) = \begin{cases} (1-\alpha)A_S(u,v) + \alpha A_T(u,v), & \text{if } u,v \in [-\hat{\beta}, \hat{\beta}] \\ A_S(u,v), & \text{otherwise} \end{cases} \quad (1)$$

where $\hat{\beta} = ||\beta H||$ or $\hat{\beta} = ||\beta W||$ for u and v respectively and $||$ is the floor rounding operation. H and W are the height and width of the image. The weighting coefficient $\alpha \in [0, 1]$ controls the mixing ratio of amplitude components from A_S and A_T. The amplitude scaling coefficient $\beta \in [0, 1]$ adjusts the area of the mutated center region and a larger β means a larger center region of A_S and A_T will be used for amplitude fusion. Eventually, with the inverse Fourier Transform F^{-1}, the reconstituted spatial-domain SD image x_S^F can be expressed as $x_S \rightarrow x_S^F = F^{-1}(A_S^F, P_S)$. Both Fourier Transform and its inverse can be efficiently implemented by the FFT [51] algorithm.

B. Chained Augmentation Mixing

Data augmentation can significantly increase generalization and robustness performance by introducing a higher diversity in training data. Furthermore, by stochastically sampling and mixing various augmentation methods with the original image, we can generate more novel augmented images without deviating too far from the original. Varieties of augmentation operations are covered in the augmentation chains (ACs), such as auto-contrast, equalization, posterization, rotation, solarization, shear, and translation in serial and parallel orientations. For a spatial-domain image x, after the chained augmentation mixing, the augmented image x^{Aug} can be expressed as below,

$$x^{Aug} = m \cdot x + (1-m) \cdot \sum_{i=1}^{AC} (w_i \cdot H_i(x)) \quad (2)$$

where m is a random convex coefficient sampled from a Beta distribution $B(\cdot)$, w_i is another random convex coefficient sampled from a Dirichlet distribution $D(\cdot)$ controlling the mixing weights of the augmentation chains, and H_i denotes

the sequential augmentation operations applied to the i^{th} augmentation chain. Each augmentation chain consists of up to three base augmentation operations that are chosen at random. Details are illustrated in the right part of Fig. 1.

C. Curri-AFDA: Curriculum-Based Augmented Fourier Domain Adaptation

For a model trained on a single domain data, it is easier to recognize images from the same domain and harder from another domain with data shift. In our curriculum strategy, the amplitude components from TD are progressively transferred to SD in the frequency space over the training period. In this way, the model learns comparatively easier information first from a single domain and successively adopts harder features like changes in the distribution of the input data from other domains. More specifically, we control the effect of the amplitude fusion by gradually increasing the amplitude scaling coefficient β from 0 to the optimal value β_{opt}. As β grows, the reconstituted training samples will gradually carry more target domain information, letting the model learn the distribution changes for the target domain. Besides, because this fusion process is slight in the early training phase, the model could firstly focus on source domain samples to recognize domain-invariant basic features without being affected by aggressive target domain information.

To facilitate our strategy that transforming the training data in each epoch following the curriculum order, i.e., "easier to harder" or "cleaner to noisier", we first employ a linearly increasing scheduler function. Specifically, if β_c is the scaling coefficient in the curriculum stage, then the linear scheduling function can be formulated as-

$$\beta_c = \begin{cases} \dfrac{e}{E \cdot r_e} \cdot \beta_{opt}, & \text{if } e \leq E \cdot r_e \\ \beta_{opt}, & \text{otherwise} \end{cases} \tag{3}$$

where E is the total number of training epochs, e is the current epoch, r_e stands for epoch ratio which controls the length of the curriculum stage in the complete training stage and further controls the changing rate of β_c with a fixed optimal scaling coefficient β_{opt}. As the epoch e grows, β_c also increases, resulting in incremental amplitude mixing as indicated in (1). In this progress, the model is gradually exposed to more target domain-specific information to improve adaptive ability continuously. Besides the linear scheduler function, there are several other candidates used in Curriculum Learning to provide distinctive learning paths. We also try the exponential scheduler function, as depicted in Fig. 2.

As a result, instead of using constant or random β in other Fourier-based adaptation methods, we adopt the incremental β_c and reconstitute the new training sample x_S^{CF} with the inverse Fourier Transform which can be represented as-

$$x_S^{CF} = F^{-1}(A^F_{S(\beta_c)}, P_S). \tag{4}$$

These generated images are then fed into the chained augmentation mixing module to produce more variations of the training data. Through this, we can improve the training data diversity further and thus boost the generalization and robustness performance of the model. The final reconstituted training image, x_S^{CAF}, can thus be expressed according to (2) as-

$$x_S^{CAF} = m \cdot x_S^{CF} + (1 - m) \cdot \sum_{i=1}^{AC} (w_i \cdot H_i(x_S^{CF})). \tag{5}$$

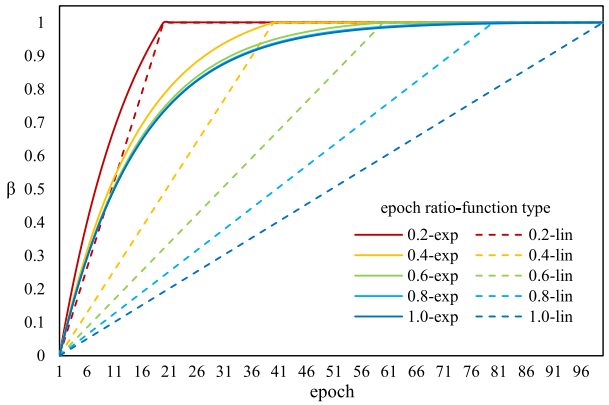


Fig. 2. **Visualization of the linear and exponential increment of amplitude scaling coefficient (β) with different epoch ratios.** Different line colors indicate different epoch ratios, and two line types differentiate two scheduling functions.

The perturbation in the low-frequency amplitude components of an image in the Fourier space will not alter the core semantics of the original image, such as the Nuclei shapes. Therefore, the masks remain unchanged in the cross-domain amplitude fusion process. Whereas in case of geometric changes during the chained augmentation mixing, the same transformations are applied to both the images and masks to adjust with the shape deviation.

Until this, we have elaborated our **Curri**culum-based **A**ugmented **F**ourier **D**omain **A**daptation (Curri-AFDA), a novel approach to resolve model degradation in case of domain shifts and data corruptions. Algorithm 1 outlines the pseudo code to implement our proposed method efficiently. In the training process, we employ Fourier Transform to acquire the amplitude components of both SD and TD images. The resultant scaling coefficient, generated by the scheduler function, regulates the amplitude fusion process. Next, the Inverse Fourier Transform is applied to produce a new image. Subsequently, the application of chained augmentation mixing facilitates the generation of additional image variants. It is worth mentioning that only one image is generated from each source-target image pair in every epoch, so there is no additional training memory consumption. However, owing to the curriculum-based cross-domain information fusion in the Fourier space, the training data distribution gradually becomes closer to the target domain. The diversity of the training data also gets boosted by the chained augmentation mixing module, which is beneficial to make the model more generalizable and robust.

IV. Experiments

A. Datasets

We perform extensive validation of our method on two widely-used and well-established medical image segmentation

Algorithm 1 Pseudo code of Curri-AFDA.

Input: Source/target domain image x_S/x_T, current/total epoch e/E, epoch ratio r_e, current/optimal amplitude scaling coefficient β_c/β_{opt}, amplitude mixing coefficient α.
Output: Final reconstituted image x_S^{CAF}.
Initialize $\alpha, \beta_{opt}, r_e, E$;
while $e < E$ **do**
  // Get Amplitude (A) and Phase (P) by Fourier Transform (F)
  $A_S, P_S \leftarrow F(x_S)$; $A_T, P_T \leftarrow F(x_T)$;
  // Scheduling scaling coefficient
  if $e \leq E \cdot r_e$ **then**
    $\beta_c \leftarrow scheduler(\beta_{opt})$;
  else
    $\beta_c \leftarrow \beta_{opt}$;
  end
  // Amplitude fusion
  $A_S^F \leftarrow AF(A_S, A_T)$ s.t. α, β_c;
  // Inverse Fourier Transform
  $x_S^{CF} \leftarrow F^{-1}(A_S^F, P_S)$;
  // Chained augmentation mixing
  $x_S^{CAF} \leftarrow CAM(x_S^{CF})$ s.t. $m \sim B(\cdot), w_i \sim D(\cdot)$;
  Take x_S^{CAF} as input for training.
end

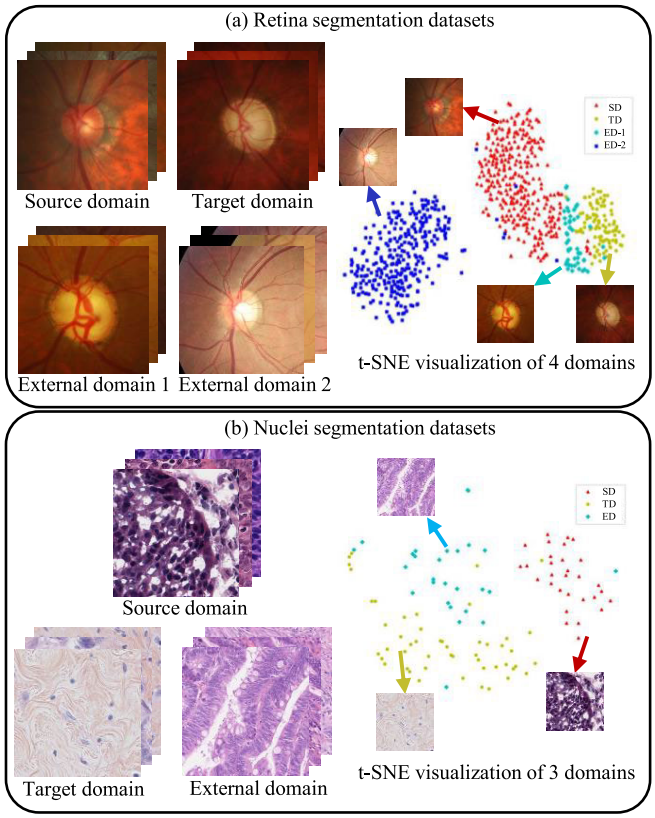


Fig. 3. **Example images of source and testing domains in (a) Retina segmentation and (b) Nuclei segmentation tasks.** The t-SNE visualization of image features (extracted by a ResNet-101 network pre-trained on ImageNet) indicates a significant domain shift.

benchmark tasks, i.e., Retina optic cup and disc segmentation on fundus images [1], [2], [52], [53] and Nuclei segmentation [11], [54], [55]. The Retina and Nuclei databases are collected from different imaging modalities, where Retina and Nuclei images are collected from color fundus photography (CFP) and pathological scanning, respectively. Besides, they comprise four and three data sources, featuring typical domain shifts such as imaging resolution, data quality, and patient populations. Therefore, they can facilitate the model assessment regarding adaptation, generalization, and robustness. For every segmentation task, we assign one domain as the source domain (SD) and select another domain as the target domain (TD) for training-time amplitude fusion. The remaining domains are considered external domains (EDs), which are unseen during training and only used to evaluate generalization and robustness performance.

Retina segmentation datasets are collected from four different scanners and sources, i.e., Drishti-GS [56], RIM-ONE-r3 [57], REFUGE-train [58], and REFUGE-valid [58]. There are two annotation labels of the optic disc and optic cup for all the datasets. These datasets are collected and pre-processed by DoFE [1] in their domain generalization task. Here we employ the database in a single-source setup containing one fixed source domain (SD), one fixed target domain (TD), and two external domains (EDs). REFUGE-train [58], RIM-ONE-r3 [57], Drishti-GS [56], and REFUGE-valid [58] have 400, 159, 101, and 400 samples, respectively. To learn more general features and mitigate potential model bias resulting from a small training data size, we designate the REFUGE-train dataset as the source domain (SD) and the remaining datasets as the target domain (TD) and two external domains (ED-1 and ED-2). Additionally, while the REFUGE-train and REFUGE-valid datasets have an identical size of 400, we maintain their original train-valid split [58] for the purposes of training and testing without any modifications. Fig. 3(a) presents some random samples from each domain and the corresponding embedded feature representation. A clear domain shift can be observed from both the appearance and the embedding space.

Nuclei segmentation datasets are collected from three sources where CryoNuSeg [59] is treated as the source domain while TNBC [60] and CoNSeP [61] are the target domain (TD) and external domain (ED). CryoNuSeg [59] dataset is extracted from the Cancer Genome Atlas (TCGA). It contains 30 images collected from 10 different human organs (three images per organ), namely the adrenal gland, larynx, lymph node, mediastinum, pancreas, pleura, skin, testis, thymus, and thyroid gland. TNBC (Triple Negative Breast Cancer) [60] dataset is acquired at Curie Institute, containing 50 images from 11 patients. The CoNSeP [61] dataset consists of 41 images, including stroma, glandular, muscular, collagen, fat and tumour regions. The data from TNBC [60] is used for cross-domain information fusion. It is randomly split for training and testing with a ratio of 8:2, similar to the dataset split strategy introduced in [62] and [63]. The testing split of TNBC [60] and the entire CoNSeP [61] dataset are not

accessed during training. The domain shift between these datasets arises from organ differences, institutional differences, and different imaging tools and protocols. Obvious domain gaps are visualized in Fig. 3(b) with the t-SNE embedding feature representations.

B. Implementation Details

We implement our method on top of a state-of-the-art segmentation backbone, UNet [64] and a recent Swin-Transformer-based model Swin-UNet [65]. A vanilla UNet architecture[1] and the official Swin-UNet implementation[2] with the pretrained Swin-Transformer [66] weight[3] are adopted. The images are resized to 384×384 and 224×224 for UNet [64] and Swin-UNet [65] models, respectively. In addition to the aforementioned Fourier-based FDA [21] and FACT [22], we also take the adversarial-based segmentation method ADVENT [13] as another baseline. We refer to the official repository[4] for implementation, such as the discriminator model and its hyper-parameters. The Fourier parts in our proposal are realized with the official implementation of FFT (Fast Fourier Transform) and IFFT (Inverse Fast Fourier Transform) from the Python Numpy library.

In the curriculum stage, the optimal amplitude scaling coefficient β_{opt} is firstly determined from the vanilla FDA [21] by empirically tuning and deriving the best constant scaling coefficient. Eventually, the value of β_{opt} is 0.006 and 1.0 in the Retina and Nuclei segmentation. Then according to (3), we schedule β_c by various epoch ratios r_e and the fixed β_{opt}. For a fair comparison, we keep the weighting coefficient α constant for all experiments. Specifically, for the UNet [64] backbone, α is set as 1.0 and 0.7 in the Retina and Nuclei Segmentation, while for the Swin-UNet [65] backbone, α is 0.5 and 0.7 in the two tasks. Further details of parameter tuning are presented and discussed in the ablation study section VI.

For the augmentation mixing, we modify the official implementation[5] of AugMix [39] to adapt it to our curriculum-based amplitude fusion training process. The augmentation level, which controls the transformation strength globally, is set as 3 and 2 in the UNet-based and Swin-UNet-based backbones. The number of augmentation chains (ACs) is 3 and each chain includes up to 3 stochastically sampled transformations. The hyperparameters in the Beta and Dirichlet distribution are all set as 1. In addition, we use a learning rate of 0.001 and the Adam optimizer for training.

To evaluate the robustness of other methods and our Curri-AFDA, we adopt various corruption techniques to construct a series of synthetic Retina datasets. Specifically, four groups of corruptions, i.e., noise, blur, weather, and digital, including 15 corruption operations, i.e., "Gaussian, Shot, Impulse", "Defocus, Glass, Motion, Zoom", "Snow, Frost, Fog, Bright", and "Contrast, Elastic, Pixel, JPEG Compression" are utilized to generate the test datasets. Furthermore, each type of corruption has five levels of severity. In this way, we can thoroughly assess the robustness under various corruption types and levels.

[1]https://github.com/ternaus/robot-surgery-segmentation
[2]https://github.com/HuCaoFighting/Swin-Unet
[3]https://github.com/microsoft/Swin-Transformer
[4]https://github.com/valeoai/ADVENT
[5]https://github.com/google-research/augmix

V. Evaluation and Results

A. Evaluation Description

To evaluate the segmentation performance, we use a commonly-used metric, Dice Similarity Coefficient (DSC). We also compute the mean and standard deviations of the results for all testing datasets to present the overall model performance. The performance of our method is compared with two closely related works, i.e., FDA [21] and FACT [22], on top of the vanilla CNN-based UNet [64] and Transformer-based Swin-UNet [65]. Besides, the GAN-based method ADVENT [13] is also adopted as another reference baseline. We have conducted extensive assessments of our method across 1) domain-adaptive performance on the target domain, 2) generalization ability on previously unseen domains, and 3) robustness to both natural and synthetic data shifts and corruptions. Our evaluation settings follow the standard unsupervised domain adaptation (UDA) [21], the generic specification of model generalization [67] and external validity [68], and the benchmark assessment of robustness [40].

In real medical applications, due to data scarcity, deep learning models are often required to handle different unseen data shifts to achieve testing-time adaptation. Considering this, we not only evaluate the training-time DA performance with TD, which is available for amplitude fusion during training, but also perform the testing-time evaluation of the generalization and robustness ability with unseen external domains (ED-1, ED-2). Data leakage is carefully considered to be avoided during training and testing. All the results reported for the testing domains are derived from testing on unseen data. Specifically, for the Retina segmentation experiments, the model is saved by considering its performance on the test-split of SD. Besides, only the train-split of TD is used for amplitude fusion during the curriculum-style training. The test split of TD and the external domains (ED-1, ED-2) are used for model evaluation. For the Nuclei segmentation task, to avoid the training and evaluation bias due to the typically small dataset size, we perform the 5-fold (folds are split based on human organs) cross-validation for training and report the average result on the left-out fold of the source domain. Similarly, the test-split of TD and the entire ED are used in performance assessment.

Regarding robustness evaluation, we test the best model derived from each method with different corruption types and levels of synthetic Retina datasets. The performance is compared in two aspects, i.e., robustness under 15 corruption types on average of 5 corruption levels and robustness under 5 corruption levels on average of 15 corruption types.

B. Results Analysis

The overall quantitative and qualitative results are shown in Table. I and Fig. 4. The results suggest the superior performance of our method Curri-AFDA in both domain-adaptive segmentation tasks compared with other methods, i.e., vanilla UNet [64], vanilla Swin-UNet [65], ADVENT [13], FDA [21] and FACT [22].

TABLE I

QUANTITATIVE RESULTS ON RETINA SEGMENTATION AND NUCLEI SEGMENTATION. FOR BOTH TASKS, ONLY THE TARGET DOMAIN (TD) IS ADOPTED FOR THE AMPLITUDE FUSION WITH THE SOURCE DOMAIN (SD) DURING TRAINING. THE EXTERNAL DOMAINS ARE USED DURING TESTING TO EVALUATE THE GENERALIZATION ROBUSTNESS. DSC (%) IS ADOPTED AS THE PERFORMANCE METRIC. AVERAGE RESULTS ACROSS ALL TESTING DOMAINS AND THE CORRESPONDING STANDARD DEVIATIONS (STD) ARE PRESENTED FOR BOTH TASKS. THE VANILLA METHOD MEANS NO DOMAIN ADAPTATION APPROACH IS APPLIED. THE BEST RESULTS ARE SHOWN IN BOLD AND THE RUNNER-UP RESULTS ARE UNDERLINED

Methods	Retina Segmentation					Nuclei Segmentation			
	β	TD	ED-1	ED-2	Mean±STD	β	TD	ED	Mean±STD
Vanilla UNet [64]	N.A.	70.33	75.52	61.65	69.17±7.01	N.A.	13.17	9.71	11.44±2.45
+ ADVENT [13]	N.A.	77.10	71.54	66.10	71.58±5.50	N.A.	27.42	28.09	27.76±0.34
+ FDA [21]	0.006	79.48	76.55	85.74	80.59±4.70	1	44.19	41.70	42.95±1.76
+ FACT [22]	1	78.18	76.29	83.00	79.16±3.46	1	33.33	32.61	32.97±0.36
+ Curri-AFDA (Ours)	linear (0 to 0.006)	78.85	**83.15**	84.59	82.20±2.99	linear (0 to 1)	46.02	46.29	46.16±0.14
	exponential (0 to 0.006)	**80.44**	80.79	**85.89**	**82.37**±3.05	exponential (0 to 1)	**50.61**	**52.29**	**51.45**±0.84
Vanilla Swin-UNet [65]	N.A.	76.89	72.57	84.33	77.93±5.95	N.A.	23.95	20.73	22.34±1.61
+ ADVENT [13]	N.A.	76.35	74.88	86.97	79.40±6.60	N.A.	40.09	**48.57**	44.33±4.24
+ FDA [21]	0.006	77.89	**77.83**	70.98	75.57±3.97	1	55.09	40.40	47.75±7.35
+ FACT [22]	1	76.74	75.36	81.55	77.88±3.25	1	52.54	45.68	49.11±3.43
+ Curri-AFDA (Ours)	linear (0 to 0.006)	**83.19**	77.81	85.24	**82.08**±3.84	linear (0 to 1)	60.62	43.03	51.83±8.80
	exponential (0 to 0.006)	78.64	73.31	**88.42**	80.12±7.66	exponential (0 to 1)	**61.75**	45.04	**53.40**±8.36

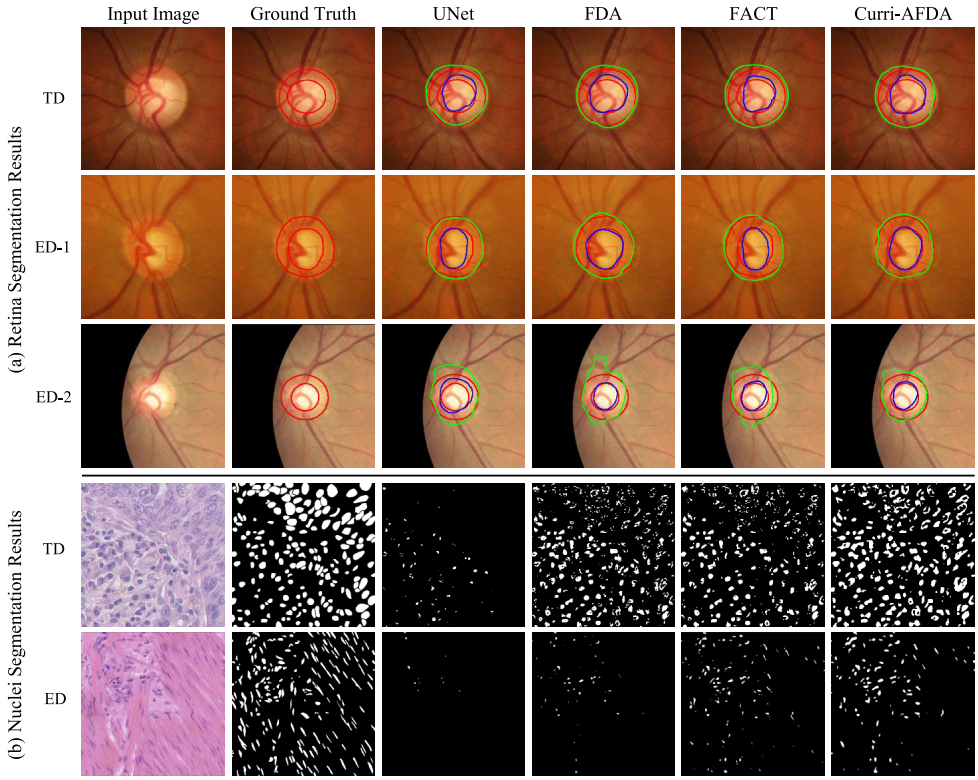


Fig. 4. **Qualitative comparison on the results of different methods with UNet [64] backbone for (a) Retina segmentation and (b) Nuclei segmentation.** Each row demonstrates the segmentation results of different methods compared with the ground truth for the testing images. In (a), the blue and green contours indicate the boundaries of the optic cups and optic discs, respectively, while the red contours are the ground truths. The boundaries of both classes obtained by our Curri-AFDA are closer to the ground truths. In (b), more nuclei can be segmented out by our method for both testing images.

1) Retina Segmentation: For the Retina Segmentation task, on the target domain (TD), our Curri-AFDA achieves the DSC improvement of 0.96% and 5.30% against the best result from the other methods with UNet [64] and Swin-UNet [65] as backbones, respectively. This shows the outstanding adaptation performance of our approach. When comparing the generalization and robustness performance on the unseen external domains, i.e., ED-1 and ED-2, our method also achieves the best result in most cases. Note that with the Swin-UNet [65] backbone, our method yields a bit lower result than the best one from other methods. We attribute this to the relatively small dataset size. However, on average of all

TABLE II

ROBUSTNESS PERFORMANCE OF OUR CURRI-AFDA AND OTHER METHODS ON CORRUPTED RETINA DATA UNDER VARIOUS TYPES OF CORRUPTION. RESULTS ARE OBTAINED BY AVERAGING THE PERFORMANCE UNDER FIVE SEVERITY LEVELS FOR EACH CORRUPTION TYPE. OUR CURRI-AFDA OUTPERFORMS OTHER METHODS BY A LARGE MARGIN FOR MOST OF THE CORRUPTION TYPES. THE BEST DSC (%) RESULTS ARE HIGHLIGHTED IN BOLD

Methods	Noise			Blur				Weather				Digital				Mean
	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG	
UNet [64]	70.77	70.24	70.49	70.05	70.09	69.19	68.39	68.43	65.45	54.78	68.78	40.19	64.49	70.28	69.74	66.09
+ FDA [21]	76.30	76.44	71.07	78.34	78.98	77.06	76.20	60.20	68.11	**64.74**	77.55	35.45	72.37	79.33	76.95	71.27
+ FACT [22]	74.55	75.08	73.30	78.03	78.44	76.17	76.21	60.20	65.27	59.95	73.53	**56.56**	71.28	78.14	75.61	71.49
+ Curri-AFDA (Ours)	**78.87**	**79.45**	**78.24**	**80.43**	**80.65**	**78.33**	**77.81**	**71.04**	**75.41**	59.89	**77.79**	52.59	**73.03**	**80.40**	**79.33**	**74.88**

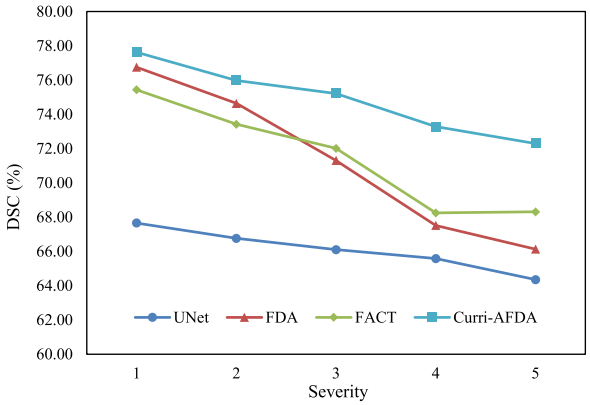


Fig. 5. **Robustness comparison of different methods on the synthetic retina data under growing severity levels of corruption on average of different corruption types.** Our Curri-AFDA is more robust to preserve higher and stabler performance.

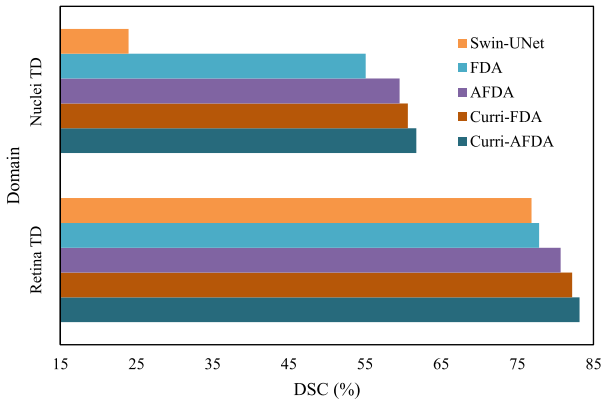


Fig. 6. **Ablation comparison of our method Curri-AFDA with the Swin-UNet [65] backbone on two target domains of Retina and Nuclei segmentation.** Consistently improved results demonstrate the efficacy of each module in our proposal.

three testing datasets, our Curri-AFDA can improve the DSC performance by 1.78% and 2.68% for the two backbones, showing the superior generalization and robustness ability. As demonstrated in Fig. 4(a), more accurate segmentation masks and boundaries can be obtained with our method.

2) Nuclei Segmentation: For the Nuclei Segmentation task which is much harder due to multiple tiny instances with uncertain positions, similar conclusions can also be drawn that our curriculum-based approach outperforms other methods regarding adaptation, generalization, and robustness, with much more significant gains. Although the performance on ED with Swin-UNet [65] is a bit lower than some other methods due to the smaller dataset size, the overall DSC gains are 8.50% and 4.29% on the testing data with UNet [64] and Swin-UNet [65] backbones. Fig. 4(b) shows the qualitative comparison of different methods for Nuclei segmentation. Our method performs better in such a segmentation task to recognize more nuclei.

3) Robustness Analysis: A more robust model is reflected in the fact that it can still maintain higher performance when exposed to corrupted images under different corruption types and increased corruption severity levels [40]. On the one hand, as shown in Table. II, our Curri-AFDA yields higher performance against most of the corruption types compared with other methods. The overall average DSC of our Curri-AFDA surpasses the best result of other methods by 3.39%. On the other hand, Fig. 5 illustrates that our Curri-AFDA maintains superior performance under increasing severity levels while the performance of other approaches degrades dramatically, especially in comparison to the other two Fourier-based approaches.

In summary, the proposed framework of curriculum-based amplitude fusion and chained augmentation mixing allows the model to explore and learn a broader feature representation space. The results of extensive experiments and evaluation on multiple domains indicate that our Curri-AFDA is generic and capable of achieving superior adaptation, generalization, and robustness performance compared with other methods.

VI. ABLATION STUDY

A. Decomposition Analysis of Each Module

As shown in Fig. 1, our proposal mainly consists of three modules, i.e., Amplitude Fusion, Curriculum Learning and Augmentation Mixing. Here we decouple these modules and compare the performance with Vanilla Swin-UNet [65] (without adaptation method), FDA [21], AFDA (FDA [21] with Augmentation Mixing) and Curri-FDA (FDA [21] with our curriculum strategy). As shown in Fig. 6, the three modules can consistently improve the performance and the integration of them, i.e., our Curri-AFDA, yields the best results on both target domains of Retina and Nuclei segmentation.

B. Curriculum Vs. Anti-Curriculum Vs. Random

Depending on a comprehensive understanding of the training data and task, the design of the curriculum is of vital significance in the utilization of Curriculum Learning. For

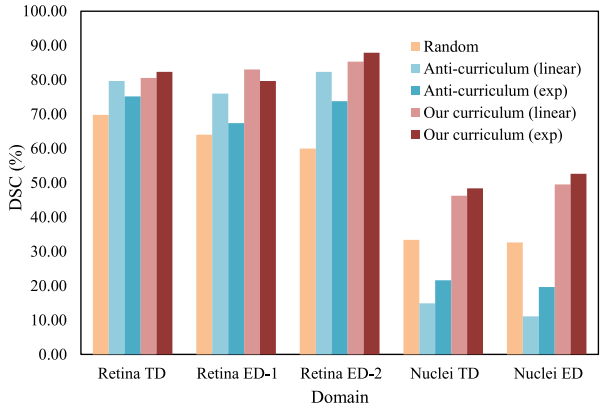


Fig. 7. **Results of different curriculum designs on testing domains of Retina segmentation and Nuclei segmentation with UNet [64] backbone.** Our curriculum can always yield better performance, especially on the nuclei datasets.

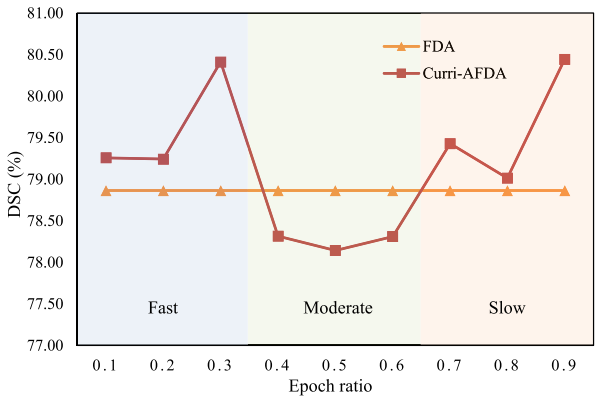


Fig. 8. **Results of tuning epoch ratios.** Faster and slower learning speeds characterized by smaller and larger epoch ratios are more likely to enhance performance.

the domain adaption and generalization task, amplitude fusion of images in the frequency space can help mitigate the variance between different domains. We take advantage of this property and establish an effective curriculum-based training framework. Specifically, in our hypothesis, the amplitude scaling coefficient β controls the amount of target domain information to be transferred to the source domain. The scheduled increment of β, i.e., the β_c in (3), is the core idea in our proposed curriculum for the domain-adaptive segmentation task. By gradually increasing the amount of the mutated low-frequency amplitude components in the source and target domain data, the generated training samples carry more domain-invariant information and thus enable the model to be more generalizable.

Apart from our curriculum, we notice that in some works [69], [70], the best curriculum is reported as the opposite of conventional curriculum learning, i.e., "harder to easier". Therefore, we also conduct experiments with another two curriculum designs, i.e., the anti-curriculum in which β_c gradually decreases and the random-curriculum in which β_c is randomly sampled in the range $[0, \beta_{opt}]$. For a comprehensive comparison, both the linear and exponential scheduling functions of β_c are evaluated and reported. As illustrated in Fig. 7, our curriculum design yields better results on all testing domains for both tasks, especially for the Nuclei segmentation task.

C. Sensitivity to Epoch Ratio

Epoch ratio (r_e in (3)) controls the duration of applying our curriculum strategy in the whole training process and affects the changing rate of the amplitude scaling coefficient β_c. This further characterizes different learning speeds of cross-domain information. We present the ablation study on the Retina segmentation task to compare the performance of our method Curri-AFDA with FDA [21] under different epoch ratios. The exponential function is adopted to update the amplitude scaling coefficient. We take the constant FDA [21] results as a reference for its irrelevance to the epoch ratio.

In our curriculum-based domain-adaptive segmentation task, smaller or larger epoch ratios result in quicker or more gradual exposure of TD information. This can let the model learn cross-domain information **faster** within earlier epochs or **slower** until later epochs. As shown in Fig. 8, these two learning speeds are more likely to enhance the model performance than a moderate one. Such behavior aligns with the generic Curriculum Leaning theory [71], which suggests that models exhibit improved performance by either learning the challenging task **faster** within earlier epochs or **slower** until later epochs, rather than adopting a moderate learning pace. The findings in this ablation study prioritize initializing the epoch ratio with a smaller or larger value to achieve optimal results.

TABLE III

COMPARISON OF LINEAR AND EXPONENTIAL SCHEDULING FUNCTIONS UNDER VARIOUS EPOCH RATIOS. THE EXPONENTIAL FUNCTION PROVIDES A HIGHER AVERAGE DSC WHILE THE LINEAR FUNCTION GIVES STABLER PERFORMANCE UNDER DIFFERENT EPOCH RATIOS WITH A LOWER STANDARD DEVIATION

Scheduler	Epoch ratio r_e								Mean	STD
	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9		
Linear	**46.02**	42.81	**45.14**	41.86	**43.79**	39.77	41.05	**42.27**	42.84	**2.08**
Exponential	38.87	**45.79**	44.40	**43.64**	41.75	**47.60**	**50.61**	40.05	**44.09**	3.91

D. Linear and Exponential Scheduling Functions

We further explore the scheduling functions of the amplitude scaling coefficient β in our curriculum. The way to update β is one of the major considerations in designing our approaches. Specifically, we implement the predefined linear and exponential functions to update β in each training epoch. As shown in Fig. 2, the key difference between them is that for a fixed epoch ratio, the exponential function yields a variable changing rate of β. In contrast, the linear function provides a constant one throughout the curriculum. Here we investigate the effect of these two scheduling functions with experiments on the Nuclei segmentation task with UNet [64] backbone.

As outlined in Section. VI-C, smaller or larger epoch ratios are more likely to yield improved results than intermediate ones. The results in Table. III further substantiate this conclusion by showing that the optimal performance of the linear and exponential schedulers are achieved with epoch ratios

TABLE IV
STATISTICS OF SKIN LESION CLASSIFICATION DATASETS. OVER 10 THOUSAND SAMPLES ARE INCLUDED IN FOUR MEDICAL DOMAINS

Domains	Lesion Types			Total
	Nevus	Benign Keratosis	Melanoma	
SD	1372	254	374	2000
TD	803	490	342	1635
ED-1	1832	475	680	2987
ED-2	3720	124	24	3868

TABLE V
QUANTITATIVE RESULTS OF SKIN LESION CLASSIFICATION WITH SWIN-TRANSFORMER [66]. THE F1 SCORE (%) RESULTS AND THE OVERALL PERFORMANCE ON THE TESTING DOMAINS ARE REPORTED. THE BEST RESULTS ARE IN BOLD AND THE RUNNER-UP RESULTS ARE UNDERLINED

Methods	β	TD	ED-1	ED-2	Mean±STD
Swin-Transformer [66]	N.A.	36.27	49.89	29.07	38.41±8.64
+ FDA [21]	0.06	40.33	49.07	39.19	42.86±4.41
+ FACT [22]	1	39.73	<u>54.01</u>	**47.64**	<u>47.13</u>±5.84
+ Curri-AFDA (Ours)	linear (0 to 0.06)	<u>41.46</u>	**56.58**	44.70	**47.58**±6.50
	exp (0 to 0.06)	**41.49**	47.69	<u>47.05</u>	45.41±2.78

of 0.2 and 0.8, respectively. These values fall within the suggested feasible range of epoch ratios. Upon examining the results presented in Table. III within the suggested ranges of epoch ratios, we can observe that the exponential scheduler outperforms the linear scheduler in more cases. Furthermore, we note that the exponential scheduler achieves a higher average performance across all epoch ratios. These findings indicate that the exponential scheduler is more reliable in providing favorable outcomes than the linear scheduler with our method.

E. Efficacy for Medical Image Classification

Besides segmentation, image classification is also fundamentally demanding in medical applications. To evaluate the efficacy of our Curri-AFDA for medical image classification, we utilize a collection of skin lesion datasets with thousands of samples released by PRR-FL [72]. The datasets have four medical domains and are annotated with three skin lesion types, i.e., Nevus, Benign Keratosis, and Melanoma. Following their dataset splits as shown in Table. IV, we conduct the classification experiments with the Swin-Transformer [66] as the backbone. The f1 score is adopted as the evaluation metric to reveal a more comprehensive comparison of the unbalanced datasets. The optimal amplitude scaling coefficient β_{opt} and the weighting coefficient α are 0.06 and 0.7, respectively.

As shown in Table. V, our method yields the best overall result of f1 score on the skin lesion datasets. This demonstrates that our methods are also supportive of the domain-adaptive medical image classification task in addition to segmentation.

VII. CONCLUSION AND DISCUSSION

This work proposes the Curriculum-based Augmented Fourier Domain Adaptation (Curri-AFDA) and proves to achieve superior adaptation, generalization, and robustness performance for medical image segmentation. Specifically, we design a novel curriculum strategy to progressively transfer amplitude information in the Fourier space from the target domain to the source domain to mitigate domain gaps and incorporate the chained augmentation mixing to further improve the generalization and robustness ability. Our method is naturally modality-independent due to its independence on any particular properties of the imaging modality. Without additional trainable parameters, extensive experiments on two segmentation tasks with multiple-domain datasets of two image modalities demonstrate the efficacy of our method on top of both the classical CNN (UNet [64]) and recent transformer (Swin-UNet [65]) architectures. Specially, we consider the crucial yet rarely explored topic in medical image analysis, i.e., the robustness performance with the synthetic dataset generated by different types and levels of corruptions, and also observe the superior results of our method. Additionally, our method can also contribute to medical image classification besides segmentation, indicating its potential for broader medical applications.

Future research may focus on designing more flexible and automatic scheduling functions to update the amplitude scaling coefficient which adjusts the amplitude fusion area. In addition, the weighting coefficient which controls the merging ratio between images can also be involved when designing the curriculum strategy. Besides the training-time Domain Adaptation, test-time Domain Adaptation [73], [74] is also worth to be explored by integrating the Fourier-based cross-domain information fusion and the chained augmentation mixing.

REFERENCES

[1] S. Wang, L. Yu, K. Li, X. Yang, C.-W. Fu, and P.-A. Heng, "DoFE: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 4237–4248, Dec. 2020.

[2] Q. Liu, C. Chen, J. Qin, Q. Dou, and P.-A. Heng, "FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 1013–1023.

[3] F. J. Piva, D. de Geus, and G. Dubbelman, "Empirical generalization study: Unsupervised domain adaptation vs. domain generalization methods for semantic segmentation in the wild," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2023, pp. 499–508.

[4] S. Rabanser, S. Günnemann, and Z. Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–13.

[5] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nature Commun.*, vol. 11, no. 1, pp. 1–10, Jul. 2020.

[6] K. Stacke, G. Eilertsen, J. Unger, and C. Lundström, "Measuring domain shift for deep learning in histopathology," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 2, pp. 325–336, Feb. 2021.

[7] R. Kocielnik, S. Kangaslahti, S. Prabhumoye, M. Hari, M. Alvarez, and A. Anandkumar, "Can you label less by using out-of-domain data? Active & transfer learning with few-shot instructions," in *Proc. Transf. Learn. Natural Lang. Process. Workshop*, 2023, pp. 22–32.

[8] Z. Cao, K. You, M. Long, J. Wang, and Q. Yang, "Learning to transfer examples for partial domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2985–2994.

[9] S. Valverde et al., "One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks," *NeuroImage, Clin.*, vol. 21, Jan. 2019, Art. no. 101638.

[10] P. Chambon, T. S. Cook, and C. P. Langlotz, "Improved fine-tuning of in-domain transformer model for inferring COVID-19 presence in multi-institutional radiology reports," *J. Digit. Imag.*, vol. 36, no. 1, pp. 164–177, Nov. 2022.

[11] M. M. Haq and J. Huang, "Adversarial domain adaptation for cell segmentation," in *Proc. Med. Imag. Deep Learn.*, 2020, pp. 277–287.

[12] M. Xu, M. Islam, C. Ming Lim, and H. Ren, "Learning domain adaptation with model calibration for surgical report generation in robotic surgery," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 12350–12356.

[13] T. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2512–2521.

[14] H. Wang and Y. Xia, "Domain-ensemble learning with cross-domain mixup for thoracic disease classification in unseen domains," *Biomed. Signal Process. Control*, vol. 81, Mar. 2023, Art. no. 104488.

[15] A. Wang, M. Islam, M. Xu, and H. Ren, "Rethinking surgical instrument segmentation: A background image can be all you need," in *Proc. 25th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Singapore: Springer, Sep. 2022, pp. 355–364.

[16] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2039–2049.

[17] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Significance-aware information bottleneck for domain adaptive semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct./Nov. 2019, pp. 6778–6787.

[18] J. Jiang et al., "PSIGAN: Joint probabilistic segmentation and image distribution matching for unpaired cross-modality adaptation-based MRI segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 4071–4084, Dec. 2020.

[19] C. Chen, Q. Dou, H. Chen, J. Qin, and P.-A. Heng, "Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation," in *Proc. 33rd AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 865–872.

[20] Y. Xia et al., "Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101766.

[21] Y. Yang and S. Soatto, "FDA: Fourier domain adaptation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4084–4094.

[22] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian, "A Fourier-based framework for domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14378–14387.

[23] Y. Yang, S. Wang, L. Zhu, P.-A. Heng, and L. Yu, "Domain generalization for medical image segmentation via hierarchical consistency regularization," 2021, *arXiv:2109.05742*.

[24] C. Yang, X. Guo, Z. Chen, and Y. Yuan, "Source free domain adaptation for medical image segmentation with Fourier style mining," *Med. Image Anal.*, vol. 79, Jul. 2022, Art. no. 102457.

[25] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.

[26] M. Islam et al., "Paced-curriculum distillation with prediction and label uncertainty for image segmentation," *Int. J. Comput. Assist. Radiol. Surgery*, pp. 1–9, Mar. 2023.

[27] E. Sangineto, M. Nabi, D. Culibrk, and N. Sebe, "Self paced deep learning for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 712–725, Mar. 2019.

[28] X. Yang, T. Burghardt, and M. Mirmehdi, "Dynamic curriculum learning for great ape detection in the wild," *Int. J. Comput. Vis.*, vol. 131, pp. 1163–1181, Jan. 2023.

[29] T. Kocmi and O. Bojar, "Curriculum learning and minibatch bucketing in neural machine translation," in *Proc. Recent Adv. Natural Lang. Process. (RANLP)*, Nov. 2017, pp. 379–386.

[30] M. Xu, M. Islam, C. M. Lim, and H. Ren, "Class-incremental domain adaptation with smoothing and calibration for surgical report generation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, Sep. 2021, pp. 269–278.

[31] M. Xu, M. Islam, B. Glocker, and H. Ren, "Confidence-aware paced-curriculum learning by label smoothing for surgical scene understanding," *IEEE Trans. Autom. Sci. Eng.*, early access, May 29, 2023, doi: 10.1109/TASE.2023.3276361.

[32] R. Portelas, C. Colas, K. Hofmann, and P.-Y. Oudeyer, "Teacher algorithms for curriculum learning of deep RL in continuously parameterized environments," in *Proc. Conf. Robot Learn.*, 2020, pp. 835–853.

[33] L. Yang, Y. Balaji, S.-N. Lim, and A. Shrivastava, "Curriculum manager for source selection in multi-source domain adaptation," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, Aug. 2020, pp. 608–624.

[34] J. Choi, M. Jeong, T. Kim, and C. Kim, "Pseudo-labeling curriculum for unsupervised domain adaptation," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, no. 67, 2019, pp. 1–13. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9474954

[35] X. Zhang, P. Shapiro, G. Kumar, P. McNamee, M. Carpuat, and K. Duh, "Curriculum learning for domain adaptation in neural machine translation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics-Hum. Lang. Technol. (NAACL-HLT)*, Jun. 2019, pp. 1903–1915.

[36] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.

[37] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "*mixup*: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–13.

[38] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6022–6031.

[39] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "AugMix: A simple method to improve robustness and uncertainty under data shift," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–15.

[40] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–16.

[41] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–22.

[42] R. Zhang, "Making convolutional networks shift-invariant again," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 7324–7334.

[43] Y. Yang, D. Lao, G. Sundaramoorthi, and S. Soatto, "Phase consistent ecological domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9008–9017.

[44] Q. Xu, R. Zhang, Z. Fan, Y. Wang, Y.-Y. Wu, and Y. Zhang, "Fourier-based augmentation with applications to domain generalization," *Pattern Recognit.*, vol. 139, Jul. 2023, Art. no. 109474.

[45] O. Stretcu, E. A. Platanios, T. M. Mitchell, and B. Póczos, "Coarse-to-fine curriculum learning," 2021, *arXiv:2106.04072*.

[46] S. Huang, X. Wang, and D. Tao, "SnapMix: Semantically proportional mixing for augmenting fine-grained data," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 2, pp. 1628–1636.

[47] H. Liu et al., "Decoupled mixup for out-of-distribution visual recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Feb. 2023, pp. 451–464.

[48] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proc. IEEE*, vol. 69, no. 5, pp. 529–541, May 1981.

[49] L. N. Piotrowski and F. W. Campbell, "A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase," *Perception*, vol. 11, no. 3, pp. 337–346, Jun. 1982.

[50] N. Guyader, A. Chauvin, C. Peyrin, J. Hérault, and C. Marendaz, "Image phase or amplitude? Rapid scene categorization is an amplitude-based process," *Comp. Rendus Biol.*, vol. 327, no. 4, pp. 313–318, Apr. 2004.

[51] H. J. Nussbaumer, "The fast Fourier transform," in *Fast Fourier Transform and Convolution Algorithms*. Berlin, Germany: Springer, 1981, pp. 80–111.

[52] Z. Zhou, L. Qi, and Y. Shi, "Generalizable medical image segmentation via random amplitude mixup and domain-specific image restoration," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*. Tel Aviv, Israel: Springer, Oct. 2022, pp. 420–436.

[53] H. Lei, W. Liu, H. Xie, B. Zhao, G. Yue, and B. Lei, "Unsupervised domain adaptation based image synthesis and feature alignment for joint optic disc and cup segmentation," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 1, pp. 90–102, Jan. 2022.

[54] Y. Sharma, S. Syed, and D. E. Brown, "MaNi: Maximizing mutual information for nuclei cross-domain unsupervised segmentation," in *Proc. 25th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Singapore: Springer, Sep. 2022, pp. 345–355.

[55] C. Li et al., "Domain adaptive nuclei instance segmentation and classification via category-aware feature alignment and pseudo-labelling," in *Proc. 25th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Singapore: Springer, Sep. 2022, pp. 715–724.

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

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

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

[59] A. Mahbod et al., "CryoNuSeg: A dataset for nuclei instance segmentation of cryosectioned H&E-stained histological images," *Comput. Biol. Med.*, vol. 132, May 2021, Art. no. 104349.

[60] P. Naylor, M. Laé, F. Reyal, and T. Walter, "Segmentation of nuclei in histopathology images by deep regression of the distance map," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 448–459, Feb. 2019.

[61] S. Graham et al., "HoVer-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101563.

[62] J. Weiss, M. Sommersperger, A. Nasseri, A. Eslami, U. Eck, and N. Navab, "Processing-aware real-time rendering for optimized tissue visualization in intraoperative 4D OCT," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2020, pp. 267–276.

[63] A. Sharghi, H. Haugerud, D. Oh, and O. Mohareri, "Automatic operating room surgical activity recognition for robot-assisted surgery," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, Sep. 2020, pp. 385–395.

[64] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Munich, Germany: Springer, Oct. 2015, pp. 234–241.

[65] H. Cao et al., "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Tel Aviv, Israel: Springer, Oct. 2023, pp. 205–218.

[66] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[67] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4396–4415, Apr. 2023.

[68] J. J. Eertink, M. W. Heymans, G. J. C. Zwezerijnen, J. M. Zijlstra, H. C. W. de Vet, and R. Boellaard, "External validation: A simulation study to compare cross-validation versus holdout or external testing to assess the performance of clinical prediction models using PET data from DLBCL patients," *EJNMMI Res.*, vol. 12, no. 1, p. 58, Sep. 2022.

[69] Y. Fan, F. Tian, T. Qin, X.-Y. Li, and T.-Y. Liu, "Learning to teach," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–16.

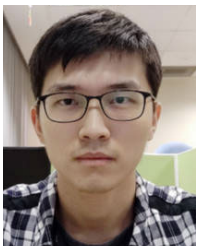
[70] W. Wang, Y. Tian, J. Ngiam, Y. Yang, I. Caswell, and Z. Parekh, "Learning a multi-domain curriculum for neural machine translation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7711–7723.

[71] X. Wu, E. Dyer, and B. Neyshabur, "When do curricula work?" in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–23.

[72] Z. Chen, M. Zhu, C. Yang, and Y. Yuan, "Personalized retrogress-resilient framework for real-world medical federated learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2021, pp. 347–356.

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

[74] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 9229–9248.



An Wang received the B.Eng. degree in information engineering from Soochow University, Suzhou, China, in 2018, and the M.Sc. degree in electrical engineering from the National University of Singapore, Singapore, in 2019. He is currently pursuing the Ph.D. degree with the Medical Mechatronics Laboratory, Department of Electronic Engineering, The Chinese University of Hong Kong, supervised by Prof. Hongliang Ren. His research interests include efficient medical image analysis and computer-assisted intervention.

Mobarakol Islam received the Ph.D. degree from the NUS Graduate School for Integrative Sciences and Engineering (NGS), National University of Singapore, in December 2019. He is currently a Senior Research Fellow with the Department of Medical Physics and Biomedical Engineering, University College London, working with Dr. Matt Clarkson with WEISS. Before that, he was a Post-Doctoral Research Associate with the Department of Computing, Imperial College London, under the supervision of Dr. Ben Glocker with the BioMedIA Laboratory. His research interests include enhancing deep neural network robustness, fairness, and reliability using calibration, uncertainty, and causality to improve image-guided disease diagnosis and intervention.

Mengya Xu received the B.Eng. degree in information engineering from Soochow University, Suzhou, China, in 2018, and the M.Sc. degree in electrical engineering from the National University of Singapore, Singapore, in 2019, where she is currently pursuing the Ph.D. degree with the Department of Biomedical Engineering, supervised by Prof. Hongliang Ren. Her research interests include vision-language multimodality-based surgical scene understanding.

Hongliang Ren (Senior Member, IEEE) received the Ph.D. degree in electronic engineering (specializing in biomedical engineering) from The Chinese University of Hong Kong (CUHK) in 2008. He has been navigating his academic journey through The Chinese University of Hong Kong, UC Berkeley, Johns Hopkins University, Children's Hospital Boston, the Harvard Medical School, the Children's National Medical Center, USA, and the National University of Singapore. He is currently an Associate Professor with the Department of Electronic Engineering, The Chinese University of Hong Kong. His research interests include robotics, mechatronics, artificial intelligence, actuators, and sensors. He has served as an active organizer and a contributor on the committees for numerous robotics conferences and delivered numerous invited keynotes/talks at flagship conferences/workshops at ICRA/IROS/ROBIO/ICIA. He was a recipient of the IFMBE/IAMBE Early Career Award in 2018, the Interstellar Early Career Investigator Award in 2018, and the ICBHI Young Investigator Award in 2019. He was a recipient of numerous international conference awards, including the Best Conference Paper Award from IEEE ROBIO 2019, IEEE RCAR 2016, IEEE CCECE 2015, IEEE Cyber 2014, and IEEE ROBIO 2013. He serves as an Associate Editor for IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING and *Medical & Biological Engineering & Computing* (MBEC).