E2EAug: An End-to-End Data Augmentation Framework for Medical Image Semantic Segmentation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Medical image semantic segmentation has numerous applications in disease detection, organ recognition, etc. Deep learning-based methods have shown remarkable efficacy in automating the segmentation of medical images. However, these methods require a large amount of annotated segmentation image-mask pairs for model training, which is a laborious, time-consuming, and costly process requiring significant domain expertise. Insufficient annotated data leads to poor generalization performance on test data. To address this problem, we propose a multi-level optimization based framework, which can train highly generalizable medical image segmentation models on a few annotated training examples. In our framework, an augmentation model and a semantic segmentation model are trained jointly, where the training of the augmentation model is guided by segmentation performance so that augmented data is tailored to improve segmentation performance. We apply our method to several medical imaging segmentation tasks. The strong generalization performance of our method is demonstrated on test images that are either in the same distribution as training data or out of the distribution of training data.

1 Introduction

Medical image semantic segmentation [1], the task of predicting the pixel-wise semantic category in a medical image, plays a crucial role in disease detection, organ recognition, tissue measurement, etc. Accurately segmenting medical images can assist medical professionals in making more precise diagnoses and planning effective treatments [2]. Various deep learning models [1, 3, 4] have been proposed for solving the semantic segmentation task. Training these models typically requires large amounts of labeled segmentation masks, which are difficult to obtain in the medical domain [5, 6], where domain-specific expertise is needed for annotating pixel-wise labels. When training data is lacking, deep learning models are prone to overfitting and generalize poorly on test data whose distribution is either the same or different from that of training data.

To mitigate the deficiency of annotated data in semantic segmentation, various methods have been proposed, including data augmentation and semi-supervised learning. Data augmentation methods [7, 8, 9, 10, 11] aim at generating synthetic image-mask pairs and using them as additional training data. For example, Neff et al. [12] leveraged WGAN-GP [13] to generate synthetic image-mask pairs from real pairs. A common limitation of these methods is that they perform data augmentation and segmentation model training separately. Segmentation performance does not guide the augmentation model during training. As a result, augmented data may not be effective for improving the segmentation model. Semi-supervised learning methods [14, 15, 16, 17, 18] leverage external unlabeled images to improve segmentation performance. For example, Li et al. [18] pre-trained a segmentation network by solving an image generation task defined on unlabeled images. These methods are limited in requiring access to large amounts of unlabeled images. In medical domains,

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

even unlabeled images are challenging to obtain due to data privacy concerns, regulation barriers (e.g., IRB approval), etc.

To address the limitations of existing methods, we propose a multi-level optimization [19] based data augmentation approach to mitigate the lack of annotated masks in semantic segmentation and improve the in-distribution and out-of-distribution generalization performance of segmentation models. Our framework consists of 1) a data augmentation model (with a learnable architecture) that augments image-mask pairs, and 2) a semantic segmentation model trained using augmented image-mask pairs. Our framework performs three learning stages end-to-end. In the first stage, we train the weight parameters of the augmentation model with its architecture tentatively fixed. In the second stage, we use the augmentation model to generate image-mask pairs, which are used to train the segmentation model. In the third stage, we evaluate the segmentation model on a validation set and update the augmentation model's architecture by minimizing validation losses. Each stage has an optimization problem. The three optimization problems are nested into a multi-level optimization problem so that they can be conducted end-to-end.

Compared with previous methods, our approach has the following advantages. First, our method trains a data augmentation model and a segmentation model end-to-end, where the segmentation model's performance guides the augmentation process of image-mask pairs. This end-to-end mechanism tailors augmented data to be effective for training a better-performing segmentation model. In contrast, previous methods perform data augmentation and segmentation model training separately, which bears a high risk that augmented data may not be suitable for improving the segmentation model. Second, unlike semi-supervised learning methods, our method does not require any unlabeled external data, which is difficult to obtain in medical domains. We evaluate our method in several medical tasks, including skin lesion segmentation, chest X-ray lung segmentation, breast cancer segmentation, placental vessel segmentation, and gastrointestinal disease segmentation. Experimental results show that our approach outperforms baselines significantly. After applying our framework for data augmentation, different segmentation models' in-distribution and out-of-distribution performance is significantly improved.

The major contributions of the paper are as follows:

- We propose an end-to-end data augmentation framework for training high-performance medical image segmentation models on limited labeled data. Without leveraging external unlabeled data, our framework can significantly improve the in-distribution and out-of-distribution generalization performance of segmentation models.
- We propose a multi-level optimization based formulation to conduct data augmentation and segmentation model training jointly.
- We apply our framework to several medical image segmentation tasks. Experimental results demonstrate the great effectiveness of our method.

2 Related works

Semantic segmentation. Many deep neural networks have been developed for semantic segmentation. Long et al. [20] proposed fully convolutional networks, which take input images of any size and generate correspondingly-sized outputs, to perform semantic segmentation. U-Net [1] aims to achieve sample-efficient semantic segmentation, where the network architecture consists of a contracting path for context capture and a symmetric expanding path for precise localization. DeepLab [3] applies atrous convolution to enlarge the view field of filters to incorporate a larger semantic context. Zhao et al. [21] proposed pyramid scene parsing network (PSPNet), which performs pyramid pooling and context aggregation to capture global context information. Yang et al. [22] proposed densely connected atrous spatial pyramid pooling (DenseASPP) for semantic segmentation, enabling the generation of multi-scale features that densely cover a larger scale range without significantly increasing the model size. Zheng et al. [23] formulated semantic segmentation as a sequence-to-sequence prediction task and used a Transformer [24] model to capture global context. Xie et al. [4] proposed SegFormer, which unifies Transformers with lightweight multilayer perceptron decoders for semantic segmentation. Our end-to-end augmentation framework is orthogonal to these segmentation models and can be applied to improve them.

Data augmentation for semantic segmentation. To mitigate the lack of annotated data in semantic segmentation, data augmentation has been investigated. Choi et al. [9] proposed an unsupervised domain adaptation method for semantic segmentation, which leverages Generative Adversarial

Networks (GANs) [25] to perform data augmentation to facilitate domain alignment. Sandfort et al. [10] employed CycleGAN [26] to transform contrast CT images into non-contrast images for data augmentation in CT segmentation tasks. Negassi et al. [11] developed Bayesian optimization based data augmentation methods for semantic segmentation. Neff et al. [12] leveraged WGAN-DP [13] to fuse a real image and its segmentation mask to create a synthetic image-mask pair. Pandey et al. [16] applied GAN to generate four-channel objects (concatenations of RGB image and segmentation mask) for nuclei image segmentation. Different from these methods which perform data augmentation and segmentation model training separately, our method performs the two tasks end-to-end.

Semi-supervised learning for semantic segmentation. Another paradigm of methods for mitigating the lack of labeled data in semantic segmentation is semi-supervised learning, which leverages unlabeled images to train segmentation models. Mendel et al. [15] proposed an Error-Correcting Supervision method for semi-supervised semantic segmentation, where a correction network predicts pseudo labels on unlabeled images. Chen et al. [17] applied consistency regularization to make the masks predicted by two segmentation models for the same unlabeled image consistent. Peng et al. [16] proposed a method that trains distinct segmentation models on various subsets of labeled data while using unlabeled images to exchange information between the models. Li et al. [18] leveraged a GAN model to learn the joint distribution of image-mask pairs from a large set of unlabeled images and a few labeled ones. Sedai et al. [27] used a teacher network to generate soft segmentation labels and uncertainty maps on unlabeled images, which are used to train a student network. Different from these methods requiring unlabeled images, our method does not rely on any external unlabeled data.

Bi-level and multi-level optimization. Many ML methods have been formulated as bi-level optimization (BLO) problems [28, 29, 30, 31, 32], where weight parameters are learned by solving a lower-level optimization problem while meta parameters are learned by solving an upper-level optimization problem. BLO-based methods have been applied for neural architecture search [33], meta learning [34], hyperparameter tuning [35], learning rate adaptation [36], data selection [37], label correction [38], etc. Following the popularity of BLO, multi-level optimization (MLO) with more than two levels of nested optimization problems has also attracted increasing attention recently [39, 40, 41, 42, 43, 44], which has been applied for data reweighting [39], data selection [40], mutual knowledge distillation [44], etc. In this paper, we leverage MLO for a new problem - end-to-end data augmentation for semantic segmentation.

3 Method

In this section, we propose an end-to-end data augmentation framework for medical image semantic segmentation based on multi-level optimization.

3.1 Overview

Our framework consists of an augmentation model and a semantic segmentation model. The augmentation model generates synthetic image-mask pairs, which are used to train the segmentation model. The validation performance of the segmentation model is leveraged to evaluate and improve the augmentation model. The two models are trained jointly to benefit from each other to achieve globally optimal performance mutually.

The augmentation model generates image-mask pairs in the following way. First, given a human-labeled segmentation mask M, some basic image augmentation operations, such as rotation, flipping, etc., are applied to M to produce an augmented mask $\widehat{M}$, which is then fed into a mask-to-image conditional GAN [45] to generate a medical image $\widehat{I}$. $(\widehat{I}, \widehat{M})$ serves as an augmented example to train the segmentation model. The conditional GAN consists of a generator and a discriminator. The generator has a learnable architecture to generate effective synthetic samples to boost performance.

Our framework consists of three learning stages, which are performed end-to-end. In the first stage, we train the weight parameters of the mask-to-image GAN. In the second stage, we use the trained GAN to generate synthetic image-mask pairs and train the segmentation model on augmented data and real data. In the third stage, we evaluate the trained segmentation model on a validation set. Validation performance reflects the fidelity of augmented data: if augmented data has low quality, the segmentation model trained using such augmentations would have poor validation performance. To improve the fidelity of augmented data, we update the architecture of the mask-to-image GAN's generator by minimizing the validation loss of the segmentation model. Each stage corresponds to an optimization problem. We integrate the three optimization problems into a nested three-level optimization formulation so that the three stages can be performed end-to-end, as shown in Figure 1.

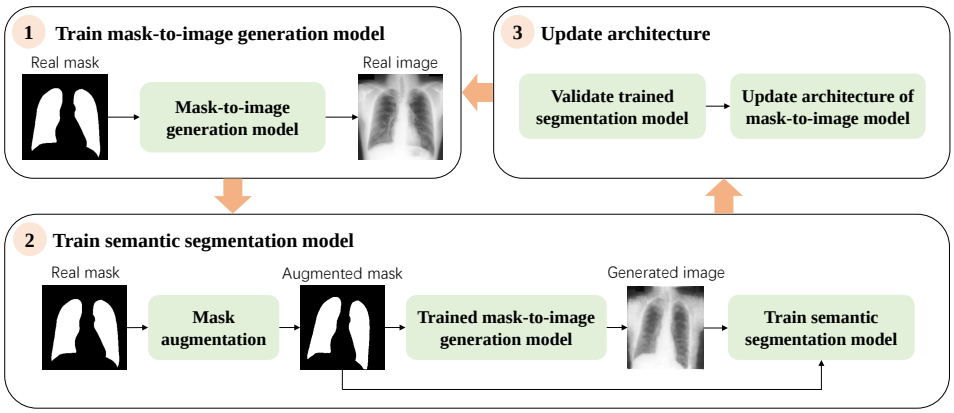

Figure 1: Overview of our framework, which consists of three stages performed end-to-end.

3.2 The end-to-end framework

Our framework consists of three learning stages, which are performed end-to-end.

Stage I. Given a medical image semantic segmentation dataset $D_{tr} = \{(I_j, M_j)\}_{j=1}^{N}$ where I_j is a real image and M_j is its segmentation mask labeled by human, we switch the order of I_j and M_j in D_{tr}, yielding $D_{gan} = \{(M_j, I_j)\}_{j=1}^{N}$ which is used to train the weight parameters G of the generator and E of the discriminator, in the mask-to-image GAN. M_j is the input of the generator and I_j is its output.

To better capture the unique properties of medical images during the generation process, we make the architecture of the generator searchable. Following [33], we perform a differentiable search, which is computationally efficient. The search space consists of a stack of computation cells. A cell is a directed acyclic graph consisting of an input node, an output node, and K operators (including convolution and transposed convolution) in the middle. Each operator is associated with a learnable selection weight $\alpha \in [0, 1]$. A larger α denotes that this operator should be selected for the final architecture. Architecture search amounts to learning these selection weights. For the i-th cell with input x_i, its output is computed as: $y_i = \sum_{k=1}^{K} \alpha_{i,k} \boldsymbol{o}_{i,k}(x_i)$, where $\boldsymbol{o}_{i,k}$ is the k-th operator of this cell and $\alpha_{i,k}$ is its selection weight. The generator's architecture is represented as $A = \{\alpha_{i,k}\}$.

This stage solves a mini-max optimization problem with a GAN-based objective function $\mathcal{L}_{gan}$:

$$G^*(A), E^* = \arg\min_G \max_E \ \mathcal{L}_{gan}\left(D_{gan}; G, A, E\right), \tag{1}$$

where the generator G aims at minimizing $\mathcal{L}_{gan}$ to generate realistic images while the discriminator E aims at maximizing $\mathcal{L}_{gan}$ to distinguish generated images from real ones. The generator's architecture A is tentatively fixed at this stage and will be updated later. The reason is that if we learn A by minimizing this training loss, a trivial solution would be yielded where A is very large and complex, which would perfectly overfit the training data. $G^*(A)$ denotes that the optimally trained generator weights G^* depends on A, since G^* depends on $\mathcal{L}_{gan}$ which depends on A.

Stage II. In the second stage, we use the mask-to-image GAN trained in the first stage to generate augmented data and use augmented data with real data to train a semantic segmentation model W. For each human-labeled segmentation mask M in D_{tr}, we apply basic image transformations to M to create an augmented mask $\widehat{M}$, which is then fed into the generator (with weights $G^*(A)$ and architecture A) to generate a medical image $f(\widehat{M}; G^*(A), A)$. $(f(\widehat{M}; G^*(A), A), \widehat{M})$ is treated as an augmented example. Let $D_{aug}(D_{tr}; G^*(A), A)$ denote the augmented dataset. We define pixel-wise cross-entropy based semantic segmentation losses $\mathcal{L}_{seg}$ on $D_{aug}(D_{tr}; G^*(A), A)$ and D_{tr}. The segmentation model W is trained by solving the following optimization problem:

$$W^*\left(G^*(A), A\right) = \arg\min_W \ \mathcal{L}_{seg}\left(D_{tr}; W\right) + \lambda \mathcal{L}_{seg}\left(D_{aug}(D_{tr}; G^*(A), A); W\right), \tag{2}$$

where λ is a tradeoff parameter, set to 1 by default.

Stage III. In the third stage, we evaluate the semantic segmentation model $W^*\left(G^*(A), A\right)$ on a validation set D_{val} consisting of real images and human-labeled masks. The generator's architecture A is updated by minimizing the validation loss:

$$\min_A \ \mathcal{L}_{seg}\left(D_{val}; W^*\left(G^*(A), A\right)\right). \tag{3}$$

Table 1: Dataset statistics

Dataset	ISIC2018	PH2	DermIS	JSRT	NLM(MC)	NLM(SZ)	BUID	FPD	KVASIR
Train	160	-	-	140	-	-	80	80	480
Valid	40	-	-	35	-	-	20	20	120
Test	594	200	98	72	138	566	230	182	200

An end-to-end framework. We perform the three stages end-to-end by integrating their optimization problems into a single multi-level optimization (MLO) problem:

$$\begin{aligned} \min_{A} \quad & \mathcal{L}_{seg}\left(D_{val}; W^*\left(G^*(A), A\right)\right) \\ s.t \quad & W^*\left(G^*(A), A\right) = \arg\min_W \ \mathcal{L}_{seg}\left(D_{tr}; W\right) + \lambda \mathcal{L}_{seg}\left(D_{aug}(D_{tr}; G^*(A), A); W\right) \\ & G^*(A), E^* = \arg\min_G \max_E \ \mathcal{L}_{gan}\left(D_{gan}; G, A, E\right) \end{aligned} \tag{4}$$

In this MLO problem, the three levels of individual optimization problems are mutually dependent. The output $G^*(A)$ of the first level is used as input to define the objective function at the second level. Similarly, the output of the second level is the input of the first level. After A is updated at the third level, the loss function at the first level changes accordingly, which renders $G^*(A)$ to change.

3.3 Optimization algorithm

We develop a gradient-based optimization algorithm for solving the problem defined in Eq.(4). Inspired by [33], we approximate $G^*(A)$ via one-step gradient descent. Then we plug the approximation into Eq.(2) and approximate $W^*(G^*(A), A)$ via one-step gradient descent similarly. Finally, we feed the approximation of $W^*(G^*(A), A)$ into Eq.(3) and update the architecture A using gradient descent. These steps iterate until convergence. Details of this algorithm are deferred to the supplements.

4 Experiments

In this section, we evaluate our proposed method on several medical image semantic segmentation tasks, including skin lesion segmentation, chest X-ray lung segmentation, breast cancer segmentation, placental vessel segmentation, and gastrointestinal disease segmentation. We evaluate both the in-distribution and out-of-distribution generalization performance of our methods.

4.1 Experimental settings

Our framework is a general and model-agnostic data augmentation method that can be applied to various semantic segmentation models. We applied our framework to two popular segmentation models, U-Net [1] and DeepLabV3 [3], by setting W in our framework to be one of them.

Datasets. For the skin lesion segmentation task, we used three datasets: ISIC2018 [46], PH2 [47], and DermIS [48]. We split ISIC2018 into training, validation, and test sets. Segmentation models were trained on the ISIC training set, and their in-distribution generalization performance was evaluated on the ISIC test set. To evaluate the out-of-distribution (OOD) generalization performance of models trained on ISIC, we tested them on PH2 and DermIS, which have different distributions than ISIC. For the chest X-ray lung segmentation task, we used three datasets: JSRT [49], NLM(MC) [50], and NLM(SZ) [50]. Segmentation models were trained on the JSRT training set, and their in-distribution generalization performance was evaluated on the JSRT test set. Their OOD generalization performance was evaluated on NLM(MC) and NLM(SZ). For breast cancer segmentation, placental vessel segmentation, and gastrointestinal disease segmentation, we used BUID [51], FPD [52], and KVASIR [53] datasets respectively to evaluate in-distribution generalization. Dataset statistics are summarized in Table 1.

Baselines. We compared our method with data augmentation baselines, including rotation, translation, flipping, their combination, and a GAN-based method [54]. Given an image-mask pair in training data, each baseline augmentation operation is applied to the input image and output mask simultaneously with the same hyperparameters, such as rotation angle, to generate augmented image-mask pairs. For the GAN-based baseline, a GAN model was trained to generate image-mask pairs. In addition, we compared with semi-supervised learning methods, including cross-teaching between CNN and Transformer (CTBCT) [55] and deep co-training (DCT) [16], which leverage unlabeled images for model training.

Metrics. For the segmentation of lung, breast cancer, and vessel, we calculated the Dice score for each patient, which is the default metric used in the literature [56] for these tasks. The Dice score is defined as $\frac{2|A \cap B|}{|A|+|B|}$, where A is the prediction and B is the ground truth. Following the ISIC challenge [57], we used the Jaccard index as the metric for skin lesion segmentation. The Jaccard index is calculated as $\frac{|A \cap B|}{|A \cup B|}$ for each patient.

Table 2: Results on skin lesion segmentation. The performance numbers are the Jaccard index (larger is better) on the test sets of ISIC, PH2, and DermIS. U-Net and DeepLab are used as backbone segmentation models. The models are trained on 40, 100, and 200 ISIC training examples, respectively. "Extra data" denotes whether external unlabeled images are used for model training.

Method	Extra data	40 ISIC training examples			100 ISIC training examples			200 ISIC training examples		
		ISIC	PH2	DermIS	ISIC	PH2	DermIS	ISIC	PH2	DermIS
U-Net [1]	✗	0.567	0.569	0.350	0.567	0.653	0.492	0.649	0.708	0.553
Rotate-UNet [7]	✗	0.587	0.640	0.421	0.647	0.722	0.606	0.641	0.740	0.657
Flip-UNet [7]	✗	0.590	0.643	0.443	0.599	0.722	0.614	0.655	0.726	0.669
Translate-UNet [7]	✗	0.570	0.656	0.406	0.620	0.730	0.620	0.651	0.728	0.675
Combine-UNet [7]	✗	0.598	0.615	0.435	0.628	0.728	0.628	0.676	0.736	0.675
GAN-UNet [54]	✗	0.612	0.710	0.583	0.633	0.743	0.632	0.681	0.752	0.677
CTBCT-UNet [55]	✓	0.661	0.724	0.632	0.604	0.702	0.566	0.665	0.743	0.667
DCT-UNet [16]	✓	0.662	0.723	0.530	0.641	0.709	0.522	0.685	0.749	0.624
Ours-UNet	✗	0.673	**0.767**	0.653	0.651	0.761	0.638	0.699	0.764	0.687
DeepLab [3]	✗	0.545	0.641	0.448	0.592	0.746	0.573	0.699	0.752	0.614
Rotate-DeepLab [7]	✗	0.603	0.667	0.577	0.602	0.701	0.630	0.705	0.787	0.646
Flip-DeepLab [7]	✗	0.628	0.694	0.618	0.674	0.719	0.605	0.726	0.783	0.675
Translate-DeepLab [7]	✗	0.584	0.679	0.599	0.605	0.766	0.574	0.730	0.778	0.646
Combine-DeepLab [7]	✗	0.637	0.705	0.624	0.594	0.748	0.647	0.742	0.791	0.684
GAN-DeepLab [54]	✗	0.642	0.729	0.633	0.684	0.759	0.655	0.749	0.807	0.693
CTBCT-DeepLab [55]	✓	0.641	0.709	0.515	0.682	0.744	0.537	0.678	0.773	0.673
DCT-DeepLab [16]	✓	0.596	0.699	0.516	0.707	0.769	0.588	0.744	0.778	0.711
Ours-DeepLab	✗	**0.695**	0.738	**0.674**	**0.736**	**0.804**	**0.675**	**0.759**	**0.814**	**0.720**

Hyperparameters. In our framework, mask augmentation is performed by sequentially applying rotation, flip, and translation, where the order of these operations in the sequence is random. Pix2Pix [54] was used as the mask-to-image model. In the architecture search space of the generator, the number of operators K in each cell was set to 3. Let Conv-xyz and UpConv-xyz denote a convolution and transposed convolution operator with kernel size x, stride y, and padding z. Candidate operators include Conv/UpConv-421, Conv/UpConv-622, and Conv/UpConv-823. In each experiment, we pre-train Pix2Pix on the training dataset. No external data was used for pre-training.

We set the number of training iterations to 5000 and chose the best checkpoint based on segmentation performance on the validation set. For training the segmentation model, we used the RMSprop optimizer [58] with an initial learning rate of $1e-5$, a momentum of 0.9, and a weight decay of $1e-3$. We used the ReduceLROnPlateau scheduler to adjust the learning rate during the training process. For training the mask-to-image GAN, we used the Adam optimizer [59] with an initial learning rate of $1e-5$, beta values of $(0.5, 0.999)$, and a weight decay of $1e-3$. For updating the architecture, we used the Adam optimizer. Experiments were conducted on A100 GPU. We ran each method three times with random initialization of model weights. We report the mean of these three runs. Their standard deviation and more implementation details are in the supplements.

4.2 Results and analysis

Skin lesion segmentation. We evaluated the generalization performance of models trained on different numbers of ISIC training examples, including 40, 100, and 200. Table 2 shows results for skin lesion segmentation. From this table, we make the following observations.

First, applying our method to U-Net and DeepLab can greatly improve their in-distribution performance. For example, when the number of training examples is 40, the Jaccard index of vanilla DeepLab on in-distribution test data ISIC is 0.55; applying our method to DeepLab (denoted as Ours-DeepLab) substantially boosts the performance to 0.70. This demonstrates that the image-mask pairs augmented by our framework are highly effective for training better segmentation models.

Second, applying our framework to U-Net and DeepLab can substantially improve their OOD generalization performance. For example, when the number of training examples is 40, applying our framework to U-Net (denoted as Ours-UNet) improves the Jaccard index on OOD test data DermIS from 0.35 to 0.65. This demonstrates that our framework is highly capable of augmenting diverse image-mask pairs, which can help the segmentation models to learn more robust and general features.

Third, to achieve similar performance as vanilla U-Net and DeepLab, our methods require much less training data. For example, with 40 training examples, Ours-DeepLab achieves a Jaccard index of 0.74 on PH2, while vanilla DeepLab needs 200 training examples to achieve similar performance. In other words, our framework improves sample efficiency by 5 times. This is more clearly visualized in

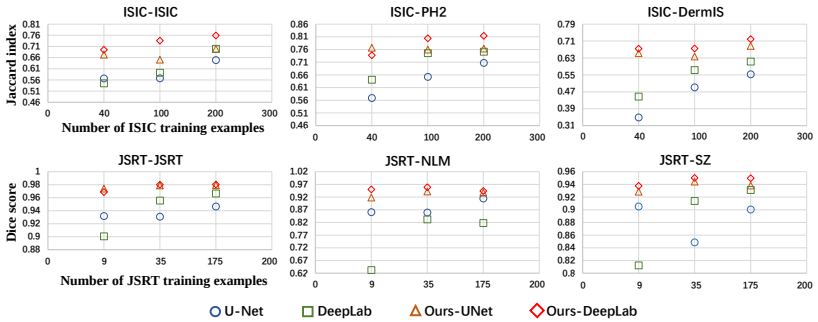

Figure 2: How performance varies with the number of training sets. In the A-B formatted title of each subfigure, A denotes a training dataset, and B denotes a test dataset. Each dot represents a method.

Figure 2, which shows how segmentation performance (y-axis) changes with the number of training examples (x-axis) for different methods. The closer to the upper left corner of each subfigure, the better a method is, in the sense that it can achieve better segmentation performance with fewer training examples. Our methods are closer to the upper left corners than baselines in all subfigures.

Fourth, our method outperforms data augmentation baselines significantly, including Rotate, Flip, Translate, Combine, and GAN-based methods. The reason is that in our multi-level optimization framework, the performance of the segmentation model guides the training of the augmentation model. Thus, augmented data is tailored to be effective for training better segmentation models. In contrast, in baseline augmentation methods, data augmentation and segmentation model training are separate. Consequently, the augmented data may not be suitable for training the segmentation models. Please see the ablation study in Figure 4 for more analysis.

Fifth, without external unlabeled images, our method outperforms semi-supervised learning (SSL) methods, CTBCT and DCT, which need external unlabeled data for training (specifically, 1000 unlabeled images in each experiment). In medical imaging applications, due to privacy concerns and regulations, even unlabeled medical images are challenging to collect, which significantly limits the applicability of SSL methods. Through leveraging unlabeled real images, SSL methods still perform worse than ours. The reason is that these methods lack the capability of inferring accurate masks for unlabeled images, i.e., they cannot construct labeled training data effectively. In contrast, our method can generate high-fidelity images from masks where the images' contents match well with the masks (see Figure 8b), effectively creating labeled training examples.

Sixth, our framework is a general and model-agnostic one that can be applied to improve different semantic segmentation models. For example, applying our framework to U-Net and DeepLab yields significant performance improvement.

Seventh, the improvement of our framework over baselines is more prominent when the number of training examples is small. This is because less training data leads to a stronger need for high-fidelity augmented data that our framework is good at generating.

Lung segmentation. The results of our method and other baselines in the lung segmentation task are shown in Table 3. Observations from this table are similar to those from Table 2. **First**, after applying our framework to different segmentation models, including U-Net and DeepLab, their in-distribution and out-of-distribution generalization performance is significantly improved. **Second**, our method outperforms data augmentation baselines, including Rotate, Flip, Translate, Combine, and the GAN-based method. **Third**, in most cases, our method performs better than semi-supervised learning baselines, including CTBCT and DCT. For these observations, the analysis of reasons is similar to that for Table 2.

Other segmentation tasks. Figure 3 shows some results for segmenting breast cancer, placental vessels, and gastrointestinal diseases on the BUID, FPD, and KVASIR datasets, respectively. As can be seen, applying our framework to U-Net and DeepLab significantly improves their performance. Due to space limits, the comparison with data augmentation and semi-supervised learning baselines is deferred to the supplements. Our method performs significantly better than these baselines.

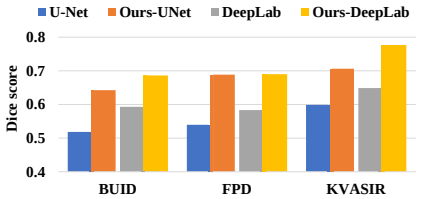

Figure 3: Results on BUID, FPD, and KVASIR datasets.

Table 3: Results on lung segmentation from chest X-rays. The performance numbers are Dice scores (larger is better) on the test sets of JSRT, MC, and SZ. Segmentation models are trained on 9, 35, and 175 JSRT training examples respectively.

Method	Extra data	9 JSRT training examples			35 JSRT training examples			175 JSRT training examples		
		JSRT	MC	SZ	JSRT	MC	SZ	JSRT	MC	SZ
U-Net [1]	✗	0.915	0.758	0.808	0.955	0.787	0.885	0.978	0.906	0.934
Rotate-UNet [7]	✗	0.914	0.781	0.826	0.965	0.852	0.891	0.979	0.933	0.934
Flip-UNet [7]	✗	0.929	0.737	0.848	0.961	0.803	0.887	0.979	0.934	0.939
Translate-UNet [7]	✗	0.923	0.779	0.826	0.962	0.811	0.883	0.978	0.922	0.941
Combine-UNet [7]	✗	0.915	0.819	0.856	0.965	0.847	0.895	0.977	0.938	0.941
GAN-UNet [54]	✗	0.921	0.816	0.849	0.964	0.849	0.896	0.979	0.939	0.933
CTBCT-UNet [55]	✓	0.926	0.843	0.864	0.954	0.813	0.899	0.969	0.929	0.938
DCT-UNet [16]	✓	0.927	0.799	0.821	0.963	0.839	0.889	0.975	0.916	0.937
Ours-UNet	✗	**0.964**	0.859	**0.928**	0.972	0.909	0.932	**0.981**	0.941	0.945
DeepLab [3]	✗	0.935	0.805	0.840	0.951	0.851	0.913	0.978	0.938	0.939
Rotate-DeepLab [7]	✗	0.943	0.846	0.848	0.954	0.891	0.915	0.979	0.947	0.939
Flip-DeepLab [7]	✗	0.935	0.844	0.843	0.953	0.882	0.921	0.978	0.946	0.941
Translate-DeepLab [7]	✗	0.935	0.846	0.859	0.958	0.896	0.916	0.969	0.940	0.944
Combine-DeepLab [7]	✗	0.946	0.854	0.866	0.961	0.897	0.914	0.978	0.933	0.936
GAN-DeepLab [54]	✗	0.952	0.845	0.889	0.966	0.898	0.925	0.978	0.944	0.940
CTBCT-DeepLab [55]	✓	0.930	0.856	0.863	0.967	0.880	0.922	0.961	0.924	**0.946**
DCT-DeepLab [16]	✓	0.946	0.859	0.847	0.957	0.882	0.905	0.959	0.939	0.940
Ours-DeepLab	✗	0.961	**0.872**	0.913	**0.975**	**0.913**	**0.942**	0.979	**0.950**	0.942

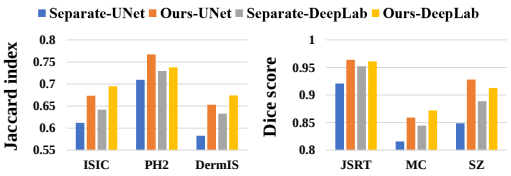

Figure 4: Ablation study results on separate training and end-to-end training.

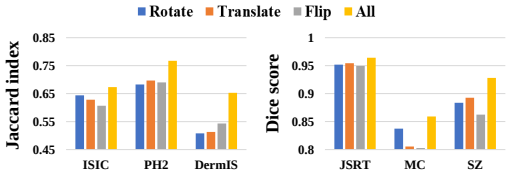

Figure 5: Ablation study results on different mask augmentation operations.

4.3 Ablation studies

To further investigate the effectiveness of individual components in our framework, we performed several ablation studies. In all experiments, models were trained on 40 ISIC examples and 9 JSRT examples. We used U-Net as the backbone segmentation model unless otherwise stated.

Ablation on end-to-end training. To further demonstrate the effectiveness of training the augmentation and segmentation models end-to-end, our end-to-end framework is compared with an ablation setting called "Separate", where the training of these two models is conducted separately. Specifically, we train the mask-to-image GAN first, fix it, and use it to generate synthetic data. Then the generated data is used to train the segmentation model. Here, we use the vanilla GAN without adjusting its architecture. The results are shown in Figure 4. As can be seen, in most cases, our end-to-end method performs much better than Separate, which further demonstrates the effectiveness of performing data augmentation and semantic segmentation jointly. In our end-to-end framework, segmentation performance closely guides the conditional GAN to generate augmented data that is effective for training the segmentation model. Such a mechanism is missing in Separate.

Sensitivity analysis of λ. In the following ablation study, we investigate how the performance of our method is affected by the hyperparameter λ in Eq.(4), which controls the tradeoff between real and augmented data when training the segmentation model. Figure 6 shows the results, where a λ value in the middle ground (e.g., 1) yields the best performance by striking the right balance between real and synthetic data.

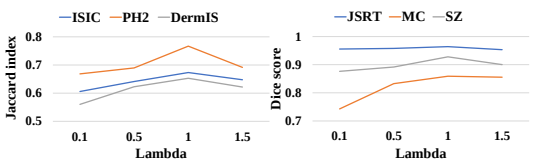

Figure 6: Ablation study results on how the tradeoff parameter λ affects segmentation performance.

Ablation on mask augmentation operations. In the third ablation study, we investigate how the choice of augmentation operations during producing augmented masks from human-labeled masks affects the final segmentation performance. We compare our full method with three ablation settings - Rotate, Translate, and Flip, which apply only a single transformation to augment masks. As shown

it looks like the generated images only move the lungs but not the body which could be a bit unrealistic

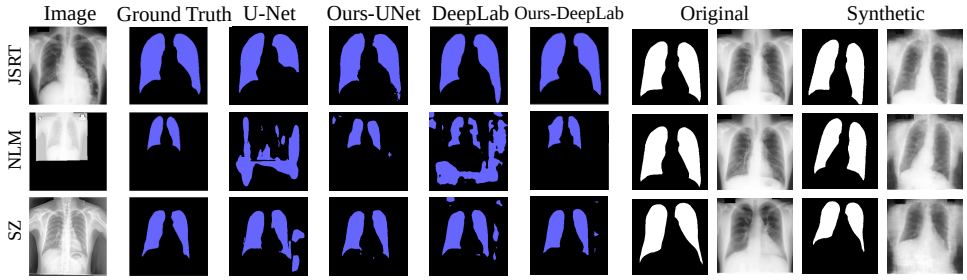

Figure 8: **(a)** Segmentation results on some randomly sampled test images. Models are trained on 175 JSRT training examples. **(b)** Examples of synthetic image-mask pairs augmented by our method.

in Figure 5, our full method, which applies three operations, outperforms the three ablation settings significantly. In particular, our framework achieves much better out-of-distribution generalization performance (on PH2, DermIS, MC, SZ). This demonstrates that compared with using a single augmentation operation, applying multiple operations is more beneficial, which can yield more diverse augmented masks and, subsequently, more diverse augmented images. Trained on diversely augmented data, segmentation models can learn robust representations and generalize better on out-of-distribution test data.

Ablation on mask-to-image GANs. Finally, we investigate how the mask-to-image conditional GAN affects the performance of the segmentation model. We compared the Pix2Pix model with SPADE [60] and ASAPNet [61], where we made their generators' architectures searchable as described in Section 3.2. Models are trained on 40 ISIC training examples with U-Net as the backbone segmentation model. Figure 7 shows the results. As can be seen, Pix2Pix and SPADE achieve similar performance while they outperform ASAPNet. The reason is that Pix2Pix and SPADE have better image generation capabilities than ASAPNet in skin lesion image generation.

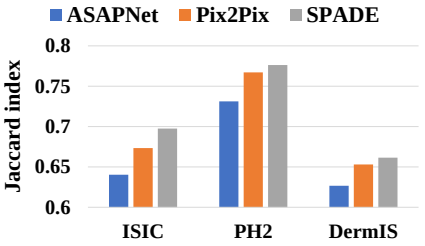

Figure 7: Ablation study results on how the mask-to-image GAN model affects segmentation performance.

4.4 Qualitative analysis

Figure 8(a) shows segmentation masks predicted by different methods on three randomly sampled chest X-ray images. The masks predicted by Ours-UNet and Ours-DeepLab are very close to the ground truth masks, while those by baselines are inaccurate. Particularly, on the NLM image, which is OOD, our methods make correct predictions while U-Net and DeepLab fail to. Figure 8(b) show some mask-image pairs augmented by Ours-UNet. As can be seen, lung regions in generated images align very well with augmented masks.

5 Conclusions and discussions

In this paper, we propose a multi-level optimization based framework to generate augmented data for medical image semantic segmentation. Our framework trains a data augmentation model and a segmentation model end-to-end, where the training of the augmentation model is guided by segmentation performance. In this way, we can tailor augmented data to be effective for training better segmentation models. Our framework consists of three learning stages, which are performed end-to-end: 1) training an augmentation model; 2) generating augmented data and training a segmentation model on augmented data; and 3) evaluating the segmentation model and updating the augmentation model's architecture by minimizing validation losses. Experiments on five segmentation tasks and various datasets demonstrate the effectiveness of our proposed method.

Limitations and broader impact. One limitation of our method is that it has a higher computational cost than simple data augmentation baselines, due to training the conditional GAN and solving the multi-level optimization problem (MLO). To address this limitation, we will leverage methods in [62, 63] to accelerate the training of GAN and employ the algorithm in [64] to improve the convergence speed of MLO. One potential negative societal impact of our work is the potential bias in the training data, which could lead to inaccurate or unfair results for patients from underrepresented communities.

References

[1] 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.

[2] S Suganyadevi, V Seethalakshmi, and K Balasamy. A review on deep learning in medical image analysis. *International Journal of Multimedia Information Retrieval*, 11(1):19–38, 2022.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[4] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[5] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021.

[6] Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21330–21340, 2022.

[7] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[8] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

[9] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6830–6840, 2019.

[10] Veit Sandfort, Ke Yan, Perry J Pickhardt, and Ronald M Summers. Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. *Scientific reports*, 9(1):16884, 2019.

[11] Misgana Negassi, Diane Wagner, and Alexander Reiterer. Smart (sampling) augment: Optimal and efficient data augmentation for semantic segmentation. *Algorithms*, 15(5):165, 2022.

[12] Thomas Neff, Christian Payer, Darko Štern, and Martin Urschler. Generative adversarial networks to synthetically augment data for deep learning based image segmentation. In *Proceedings of the OAGM Workshop*, pages 22–29, 2018.

[13] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

[14] Man Zhang, Yong Zhou, Jiaqi Zhao, Yiyun Man, Bing Liu, and Rui Yao. A survey of semi-and weakly supervised semantic segmentation of images. *Artificial Intelligence Review*, 53:4259–4288, 2020.

[15] Robert Mendel, Luis Antonio De Souza, David Rauber, Joao Paulo Papa, and Christoph Palm. Semi-supervised segmentation based on error-correcting supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 141–157. Springer, 2020.

[16] Jizong Peng, Guillermo Estrada, Marco Pedersoli, and Christian Desrosiers. Deep co-training for semi-supervised image segmentation. *Pattern Recognition*, 107:107269, 2020.

[17] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021.

[18] 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.

[19] Sang Keun Choe, Willie Neiswanger, Pengtao Xie, and Eric Xing. Betty: An automatic differentiation library for multilevel optimization. *arXiv preprint arXiv:2207.02849*, 2022.

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[21] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[22] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3684–3692, 2018.

[23] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[26] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[27] Suman Sedai, Bhavna Antony, Ravneet Rai, Katie Jones, Hiroshi Ishikawa, Joel Schuman, Wollstein Gadi, and Rahil Garnavi. Uncertainty guided semi-supervised segmentation of retinal layers in oct images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, pages 282–290. Springer, 2019.

[28] Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

[29] Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pages 3748–3758. PMLR, 2020.

[30] Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, pages 4882–4892. PMLR, 2021.

[31] Risheng Liu, Yaohua Liu, Shangzhi Zeng, and Jin Zhang. Towards gradient-based bilevel optimization with non-convex followers and beyond. *Advances in Neural Information Processing Systems*, 34, 2021.

[32] Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

[33] Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: differentiable architecture search. In *ICLR*, 2019.

[34] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

[35] Matthias Feurer, Jost Springenberg, and Frank Hutter. Initializing bayesian hyperparameter optimization via meta-learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), 2015.

[36] Atilim Gunes Baydin, Robert Cornish, David Martinez Rubio, Mark Schmidt, and Frank Wood. Online learning rate adaptation with hypergradient descent. In *International Conference on Learning Representations*, 2018.

[37] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems*, pages 1919–1930, 2019.

[38] Guoqing Zheng, Ahmed Hassan Awadallah, and Susan Dumais. Meta label correction for noisy label learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11053–11061, 2021.

[39] Bhanu Garg, Li Zhang, Pradyumna Sridhara, Ramtin Hosseini, Eric Xing, and Pengtao Xie. Learning from mistakes-a framework for neural architecture search. *AAAI Conference on Artificial Intelligence*, 2022.

[40] Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Towards visual question answering on pathology images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 708–718, 2021.

[41] Aniruddh Raghu, Jonathan Lorraine, Simon Kornblith, Matthew McDermott, and David K Duvenaud. Meta-learning to improve pre-training. *Advances in Neural Information Processing Systems*, 34, 2021.

[42] Sai Ashish Somayajula, Linfeng Song, and Pengtao Xie. A multi-level optimization framework for end-to-end text augmentation. *Transactions of the Association for Computational Linguistics*, 10:343–358, 2022.

[43] Felipe Petroski Such, Aditya Rawal, Joel Lehman, Kenneth Stanley, and Jeffrey Clune. Generative teaching networks: Accelerating neural architecture search by learning to generate synthetic training data. In *International Conference on Machine Learning*, pages 9206–9216. PMLR, 2020.

[44] Pengtao Xie and Xuefeng Du. Performance-aware mutual knowledge distillation for improving neural architecture search. *CVPR*, 2022.

[45] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[46] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[47] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. Ph 2-a dermoscopic image database for research and benchmarking. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 5437–5440. IEEE, 2013.

[48] Jeffrey Luc Glaister. Automatic segmentation of skin lesions from dermatological photographs. Master's thesis, University of Waterloo, 2013.

[49] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1):71–74, 2000.

[50] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

[51] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

[52] Sophia Bano, Francisco Vasconcelos, Luke M. Shepherd, Emmanue Vander Poorten, Tom Vercauteren, Sebastien Ourselin, Anna L. David, Jan Deprest, and Danail Stoyanov. Deep placental vessel segmentation for fetoscopic mosaicking. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020.

[53] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, pages 451–462. Springer, 2020.

[54] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[55] Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In *International Conference on Medical Imaging with Deep Learning*, pages 820–833. PMLR, 2022.

[56] Jeroen Bertels, Tom Eelbode, Maxim Berman, Dirk Vandermeulen, Frederik Maes, Raf Bisschops, and Matthew B Blaschko. Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 92–100. Springer, 2019.

[57] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):34, 2021.

[58] Naichen Shi and Dawei Li. Rmsprop converges with proper hyperparameter. In *International conference on learning representation*, 2021.

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

[60] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[61] Tamar Rott Shaham, Michaël Gharbi, Richard Zhang, Eli Shechtman, and Tomer Michaeli. Spatially-adaptive pixelwise networks for fast image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14882–14891, 2021.

[62] Samarth Sinha, Han Zhang, Anirudh Goyal, Yoshua Bengio, Hugo Larochelle, and Augustus Odena. Small-gan: Speeding up gan training using core-sets. In *International Conference on Machine Learning*, pages 9005–9015. PMLR, 2020.

[63] Samarth Sinha, Zhengli Zhao, Anirudh Goyal ALIAS PARTH GOYAL, Colin A Raffel, and Augustus Odena. Top-k training of gans: Improving gan performance by throwing away bad samples. *Advances in Neural Information Processing Systems*, 33:14638–14649, 2020.

[64] Ryo Sato, Mirai Tanaka, and Akiko Takeda. A gradient method for multilevel optimization. *Advances in Neural Information Processing Systems*, 34:7522–7533, 2021.