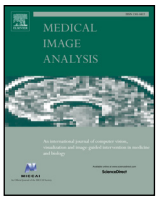
Medical Image Analysis 78 (2022) 102394

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media



Non-equivalent images and pixels: Confidence-aware resampling with meta-learning mixup for polyp segmentation

Xiaoqing Guo[a], Zhen Chen[a], Jun Liu[c], Yixuan Yuan[a,b,*]

[a] Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR, China
[b] City University of Hong Kong Shenzhen Research Institute, Shenzhen, PR China
[c] Department of Mechanical Engineering, City University of Hong Kong, Hong Kong SAR, China

ARTICLE INFO

Article history:
Received 19 August 2021
Revised 7 December 2021
Accepted 11 February 2022
Available online 18 February 2022

Keywords:
Meta-learning mixup
Confidence-aware resampling strategy
Polyp segmentation

ABSTRACT

Automatic segmentation of polyp regions in endoscope images is essential for the early diagnosis and surgical planning of colorectal cancer. Recently, deep learning-based approaches have achieved remarkable progress for polyp segmentation, but they are at the expense of laborious large-scale pixel-wise annotations. In addition, these models treat samples equally, which may cause unstable training due to polyp variability. To address these issues, we propose a novel Meta-Learning Mixup (MLMix) data augmentation method and a Confidence-Aware Resampling (CAR) strategy for polyp segmentation. MLMix adaptively learns the interpolation policy for mixup data in a data-driven way, thereby transferring the original soft mixup label to a reliable hard label and enriching the limited training dataset. Considering the difficulty of polyp image variability in segmentation, the CAR strategy is proposed to progressively select relatively confident images and pixels to facilitate the representation ability of model and ensure the stability of the training procedure. Moreover, the CAR strategy leverages class distribution prior knowledge and assigns different penalty coefficients for polyp and normal classes to rebalance the selected data distribution. The effectiveness of the proposed MLMix data augmentation method and CAR strategy is demonstrated through comprehensive experiments, and our proposed model achieves state-of-the-art performance with 87.450% dice on the EndoScene test set and 86.453% dice on the wireless capsule endoscopy (WCE) polyp dataset.

1. Introduction

Colorectal cancer (CRC) is the second most common cause of cancer-related deaths in the United States, with 52,980 estimated deaths in 2021 (Siegel et al., 2021). Fortunately, if adenomatous polyps, i.e., precursors to CRC, are detected and removed before they develop into malignant tumors, deaths caused by CRC can be significantly reduced, with a favorable 5-year survival rate of 90% (Siegel et al., 2021). Colonoscopy and WCE are common diagnostic tools that are used in regular screening procedures to identify the adenomatous polyps (Jia et al., 2019). This procedure is usually performed manually by clinicians and can be subjected to human errors and missed diagnosis of polyps. Hence, an automatic and reliable polyp region segmentation model is highly demanded for assisting clinicians in the diagnostic process.

In the last decades, numerous deep learning-based convolutional neural networks (CNN) have been developed for the automatic polyp detection and segmentation (Vázquez et al., 2017; Zhou et al., 2018; Akbari et al., 2018; Yuan et al., 2018; Guo and Yuan, 2019; Jha et al., 2019; Fang et al., 2019; Qadir et al., 2019; Wickstrøm et al., 2020; Zhang et al., 2020a; Jia et al., 2020; Fan et al., 2020; Nguyen et al., 2020; Yang et al., 2020; Lin et al., 2020; Wu et al., 2021; Liu et al., 2021; Chen et al., 2021; Jha et al., 2021; Guo et al., 2021a; Yang et al., 2021; Guo et al., 2021b). Most of these methods are based on encoder-decoder network architectures, where polyp segmentation masks are learned in an end-to-end manner and supervised by pixel-wise annotations. The success of deep CNNs usually depends on the sufficient annotated data. Despite considerable progresses, current polyp segmentation algorithms still could not fulfill clinical requirements (Wu et al., 2021), and the design of automatic polyp recognition system with desirable reliability remains challenging due to *the lack of abundant annotated datasets* and *the variability of obtain polyp images*.

The lack of abundant annotated datasets is the first obstacle. High-quality annotated datasets are scarce in polyp segmentation

* Corresponding author at: Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR, China.
E-mail address: yxyuan.ee@cityu.edu.hk (Y. Yuan).

https://doi.org/10.1016/j.media.2022.102394
1361-8415/

tasks, because the acquisition of expert pixel-level annotations requires a high degree of attentiveness and domain expertise. A feasible solution is to enlarge the limited dataset via data augmentation (Krizhevsky et al., 2012; Vázquez et al., 2017; Guo et al., 2021a). However, traditional augmentation methods, such as random rotation, only create new images that are similar to original training images. Although the recently proposed mixup generates noticeably different training images by implementing an pixel-wise linear combination of data (Zhang et al., 2018), it is designed for image classification, where soft label is produced for mixed image. Directly applying it to polyp segmentation is problematic, because the soft label calculated in the vanilla mixup method may yield inconsistent category information with the clinical diagnosis, which degrades the segmentation performance. Inspired by the meta-learning strategy, which has successfully shown to be powerful in learning data-driven policies using the knowledge of validation set (meta-data) (Finn et al., 2017; Wang et al., 2020), we propose Meta-Learning Mixup (MLMix) data augmentation method to learn the data-driven interpolation policy of mixed label and generate compatible hard label for mixed image. Specifically, the data-driven interpolation policy transfers the soft label to an optimal hard label and targets to improve the accuracy on the validation set. Since the validation set is constructed by manually annotated data, it provides meta-knowledge for interpolation policy learning and guarantees the consistent category information with the clinical diagnosis. Through the meta-learning strategy, the proposed MLMix has capability of obtaining the matched segmentation label for each augmented mixup image and facilitates the generalization ability of segmentation model.

The second challenge lies in that existing polyp segmentation models treat samples equally in the loss calculation, e.g., cross-entropy loss, ignoring the negative effect caused by *the variability of polyps* (Guo et al., 2021b) in endoscope images. These images are obtained from different patients, and polyps can locate at any spots of the image with different illumination situations and partial obstructions. Such variability prevents neural networks from effectively learning the general patterns of polyps and causes them to get stuck at suboptimal solutions (Li and Gong, 2017). Thus, it may be beneficial to first train networks on more representative polyp images and regions, and then gradually introduce more challenging instances. Although the easy-to-hard hierarchical learning strategy has been considered by previous studies (Li and Gong, 2017; Qin et al., 2020) to prevent the fluctuation of training procedure, they only select confident samples in image level and ignore the class imbalance problem, which is common in polyp segmentation. In practice, only less than 10% pixels belong to the polyp category (Guo et al., 2021a), and the imbalanced training data cause the polyp category to be heavily under-represented. In contrast to previous studies (Li and Gong, 2017; Qin et al., 2020), we recognize that the image- and pixel-level easy-to-hard gradual learning schemes address the polyp variability and class imbalance problems from complementary perspectives, i.e., pixels with high confidence scores in those unselected complex images also play an important role in model optimization. Therefore, we consider that jointly performing both easy-to-hard gradual learning aspects could impart their individual advantages and subsequently improve the segmentation performance, and thus, we devise a Confidence-Aware Resampling (CAR) strategy for polyp segmentation. Specifically, images are firstly ranked according to the image-level confidence (overlap scores with ground truth, horizontal axis in Fig. 1) in ascending order. Considering there exist informative pixels in those unselected complex images, we also rank pixels according to the pixel-level confidence (classification accuracy, vertical axis in Fig. 1). The proposed CAR strategy adaptively selects confident images (Fig. 1(a)) and pixels (Fig. 1(b, c)) to gradually mine informative samples and enable robust training. Moreover, the proportions of selected foreground and background pixels are assigned individually based on the class distribution prior knowledge to ensure the balance of the selected class distribution.

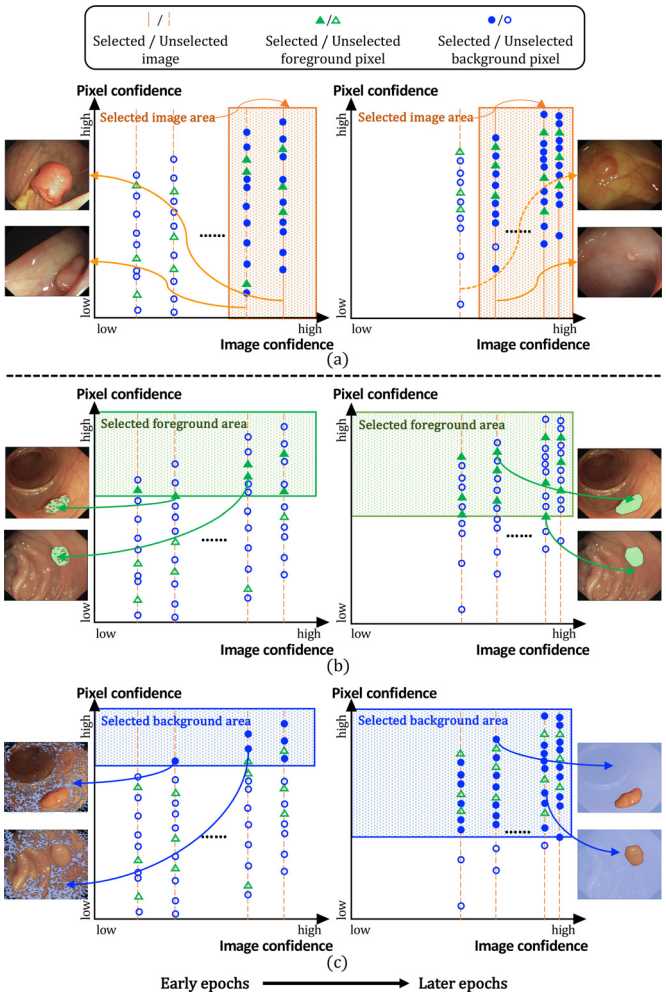


Fig. 1. Motivation of the proposed CAR strategy. In order to ensure the stable training of the segmentation net, we implement (a) the image-level selection and (b, c) the pixel-level selection in an easy-to-hard gradual learning manner for model optimization. Note that selected (b) foreground pixels (green regions) and (c) background ones (blue regions) are rebalanced by assigning different penalty proportions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Our main contributions can be summarized as follows:

- We propose a novel MLMix data augmentation method to enlarge the limited annotated training dataset for promoting the generalization ability of the optimized segmentation model. To our best knowledge, the proposed MLMix represents the first effort that leverages meta-learning strategy to gain clinical diagnosis knowledge from validation data, so as to generate compatible hard label for mixed image in a data-driven manner.
- We further develop a CAR strategy to enable the robust training of polyp segmentation model and accelerate the learning process, which adopts the easy-to-hard gradual learning scheme in both image and pixel levels. Moreover, the class prior knowledge is integrated in the training procedure to rebalance the data distribution for mitigating the class imbalance problem.
- The effectiveness and generality of MLMix and CAR strategy are validated on the EndoScene dataset (Vázquez et al., 2017) and WCE polyp dataset. Extensive experiments show that our approach outperforms state-of-the-art polyp segmentation methods.

The rest of the paper is organized as follows: Section 2 reviews the related work of polyp segmentation, mixup and resampling/reweighting strategy. Then the proposed methods including MLMix and CAR strategy are illuminated in Section 3. The experimental results are analyzed in Section 4, and we conclude in Section 5.

2. Related work

2.1. Deep learning for polyp segmentation

Owing to the excellent feature representation capacity, deep learning methods based on CNNs have been widely employed for automatic polyp detection and segmentation (Vázquez et al., 2017; Zhou et al., 2018; Akbari et al., 2018; Yuan et al., 2018; Guo and Yuan, 2019; Jha et al., 2019; Fang et al., 2019; Qadir et al., 2019; Wickstrøm et al., 2020; Zhang et al., 2020a; Jia et al., 2020; Fan et al., 2020; Nguyen et al., 2020; Yang et al., 2020; Lin et al., 2020; Wu et al., 2021; Liu et al., 2021; Chen et al., 2021; Jha et al., 2021; Guo et al., 2021a; Yang et al., 2021; Guo et al., 2021b). Vázquez et al. (2017) made the first attempt to integrate the deep learning algorithm in polyp segmentation by utilizing a fully convolutional network (FCN), which enables deep neural network to make spatially dense predictions. Qadir et al. (2019) introduced Mask R-CNN to simultaneously perform polyp detection and segmentation. In order to prevent fragmentary predictions, Jia et al. (2020) extended FCN by introducing a region proposal stage. Wickstrøm et al. (2020) leveraged the pooling indices derived in the max-pooling operator of an encoder to implement non-linear upsampling in the decoder, thus preserving the spatial dependence. To further diminish the semantic discrepancies between deep and shallow layers, Zhou et al. (2018) designed the UNet++ network architecture with nested skip connections among different layers. Furthermore, Fang et al. (2019) introduced a boundary constraint to make the segmentation model more sensitive to predictions around polyp boundaries. Recently, increasing number of studies are attempting to aggregate effective information for polyp segmentation (Zhang et al., 2020a; Fan et al., 2020). For example, Zhang et al. (2020a) designed a context selection based segmentation framework to dynamically combine global and local contextual information for regional contrast reasoning. Fan et al. (2020) proposed a parallel reverse attention network to recurrently mine the relationship between polyp area and boundaries.

Although these polyp segmentation models have made significant progress, due to the limited annotation data and huge polyp variability, they may still perform unsatisfactorily in clinical applications.

2.2. Mixup

Correctly delineating polyp regions is challenging for clinicians due to the various shapes, textures, and illumination situations exhibited in endoscope images, and even professional clinicians may yield discrepant annotation results. Hence, high-quality annotated datasets are scarce in polyp segmentation tasks. Fortunately, data augmentation provides a convenient way for enriching the training samples and can significantly facilitate the generalization ability of deep CNN models (Zhang et al., 2018; Li et al., 2019; Chaitanya et al., 2019; Wang et al., 2019; Berthelot et al., 2019; Verma et al., 2019; Hendrycks et al., 2020; Guo et al., 2021a). The recently proposed mixup performs convex combinations on images and labels to produce noticeably dissimilar data to original image (Zhang et al., 2018). Inheriting its excellent properties, increasing interest has been paid to employing and amending mixup in diverse applications (Li et al., 2019; Wang et al., 2019; Berthelot et al., 2019; Verma et al., 2019). Li et al. (2019) proposed an asymmetric mixup that could explicitly keep the decision boundary of classifier close to the majority category and stay away from the minority category, thus alleviating the class imbalance problem. To reduce distribution mismatch, Berthelot et al. (2019) assigned low-entropy pseudo labels for unlabeled examples and then implemented mixup between labeled and unlabeled images. Moreover, Verma et al. (2019) devised manifold mixup to exploit the interpolation at hidden representations, thereby optimizing neural networks with smoother decision boundaries at different feature levels.

However, directly employing the mixup algorithm, which was designed for whole-image classifications, to the polyp segmentation task is problematic, because mixup produces soft label and ignores the varying degrees of CRC. Our previous study (Guo et al., 2021a) tackled this issue by devising a confidence-guided manifold mixup to enrich training data in both image and feature levels. Herein, we present a novel alternative, coined MLMix, to generate compatible hard segmentation labels for mixed endoscope images in a data-driven manner.

2.3. Resampling or reweighting strategy

To address the biased data problems, such as the variability of samples and class imbalanced data distribution, resampling and reweighting strategies have been well studied in the literature (Li and Gong, 2017; Lin et al., 2017; Jiang et al., 2018; Li et al., 2019; Li and Vasconcelos, 2020; Qin et al., 2020; Cai et al., 2020). On the one hand, self-paced learning (SPL) progressively incorporates increasing number of images in an easy-to-hard manner to enable a robust model that learned with sample variability (Li and Gong, 2017). Qin et al. (2020) measured the prostate segmentation difficulty for all images and gradually selected the relatively confident samples for segmentation model optimization. Recently, authors in (Jiang et al., 2018; Cai et al., 2020) utilized the multiple-layer perception to automatically assign a large weighting coefficient for an easy sample. The strategy of prioritizing training on confident samples (Li and Gong, 2017; Jiang et al., 2018; Cai et al., 2020; Qin et al., 2020) exhibited superior performance in real problems involving sample variability and noisy labeled data. On the other hand, some methods (Lin et al., 2017; Li et al., 2019; Li and Vasconcelos, 2020) emphasized samples with large loss values to mitigate the class imbalance problem, because they can adaptively assign low weights for the majority class instances and penalize the minority class instances with relatively high loss values, thereby ensuring a balanced loss calculation.

Nevertheless, these methods cannot simultaneously deal with the variability of samples and imbalanced data distribution in polyp segmentation. In contrast, the proposed CAR strategy adopts an easy-to-hard gradual learning scheme and leverages class prior knowledge to tackle the aforementioned two biased data problems.

3. Method

Fig. 2 illustrates the overall framework of the proposed model. First, the proposed MLMix (§3.1) is utilized to augment the original dataset $\mathcal{D} = (\mathcal{X}, \mathcal{Y})$ as $\widetilde{\mathcal{D}} = (\widetilde{\mathcal{X}}, \widetilde{\mathcal{Y}})$, enriching the limited training data. Then, the original and augmented datasets are combined and fed into the segmentation net, which is denoted as $f_{\mathbf{w}}(\cdot)$ and parameterized by $\mathbf{w}$. The proposed CAR strategy (§3.2) progressively selects the confident image set $\mathbb{S}_{img}$ with higher overlap scores and the class balanced pixel set $\mathbb{S}_{pix}$ with higher classification confidence scores, and it simultaneously utilizes them together to retrain the segmentation net. With the easy-to-hard training strategy, the segmentation net can accelerate the learning convergence,

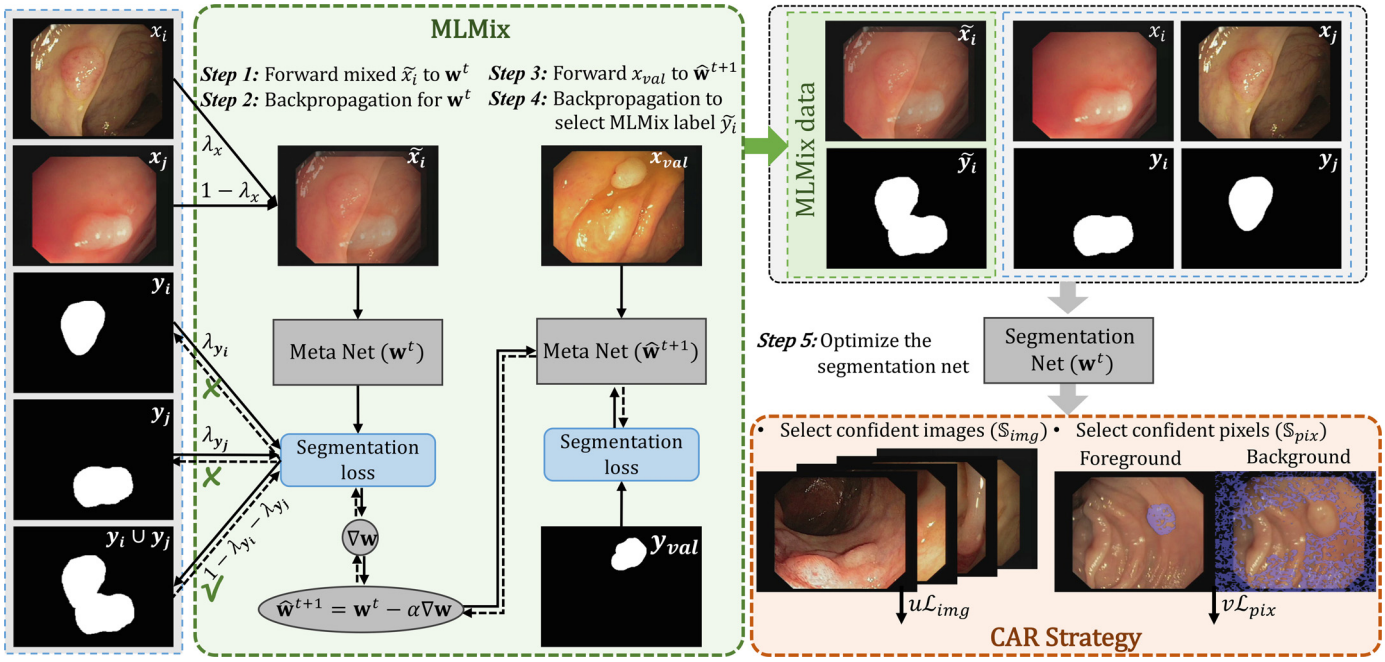


Fig. 2. Illustration of the proposed segmentation model with Meta-Learning Mixup (MLMix) and Confidence-Aware Resampling (CAR) strategy. Step 1–4: meta steps to obtain the compatible hard label $\widetilde{y}_i$ for mixed image $\widetilde{x}_i$. Step 5: optimization of segmentation net with CAR strategy on MLMix data and original data.

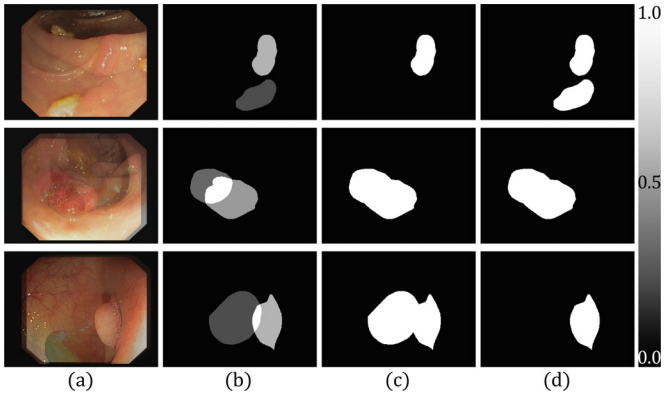


Fig. 3. Each column presents (a) mixup image, (b) mixup label (Zhang et al., 2018), (c) asymmetric mixup label (Li et al., 2019), (d) MLMix label.

improve the generalization capability, and rebalance the data distribution. During the inference phase, test images are directly fed into the segmentation net to obtain their predictions.

3.1. Meta-learning mixup (MLMix)

Mixup is an effective data augmentation algorithm that can promote the model generalization ability for image classification task. It generates extra training samples through convex combinations (Zhang et al., 2018; Guo et al., 2021a). Given randomly sampled image-label pairs (x_i, y_i) and (x_j, y_j) from the training data, the augmented mixup image and the corresponding mixup label are computed via $\widetilde{x}_i = \lambda x_i + (1-\lambda)x_j$ and $\widetilde{y}_i = \lambda y_i + (1-\lambda)y_j$, where λ is drawn from a beta distribution. Directly employing the mixup, which was designed for image classification, to the polyp image segmentation model is problematic, because it produces soft labels and provides inconsistent category information with clinical diagnosis, as shown in Fig. 3 (b). In the clinical diagnosis, appearances of polyps are evidently different for varying degrees of CRC. Although mixing malignant polyp regions with normal ones will weaken the feature representation of polyp and obscure the lesion boundaries (just like polyp at its early stage), the corresponding area should be diagnosed as a polyp in clinical. To remedy this inconsistent diagnosis issue, asymmetric mixup (Li et al., 2019) introduces a margin coefficient m to threshold the mixup label, and the modified hard label is denoted as follow:

$$\widetilde{y}_i = \delta(\lambda y_i + (1-\lambda)y_j > m), \tag{1}$$

where $\delta(\cdot) = 1$ if the condition is fulfilled, and otherwise $\delta(\cdot) = 0$. The mixed region, with a label value above a certain margin m, should be classified into the polyp category. The Fig. 3 (c) column represents the asymmetric label with $m = 0.3$, and it is observed that applying a uniform margin coefficient may yield inaccurate hard labels, as illustrated in the upper and lower rows. Hence, learning a data-driven hard label for polyp segmentation will be beneficial. To this end, we incorporate the meta-learning strategy and propose MLMix, which generates compatible label for mixed image in an online fashion. Our intuition is that the meta-learning strategy learning to gain clinical diagnosis knowledge from validation data (i.e., manually annotated data) can provide instructive supervision for refining the interpolation policy of mixed label in a data-driven manner.

Considering the characteristics of unclear polyps may be overwhelmed by normal regions and the malignant polyp mixed with normal tissues may still belong to polyp, we first decouple the hard label of $\widetilde{y}_i$ in Eq. (1) into $\{y_i, y_j, y_i \cup y_j\}$ to comprehensively cover all cases of hard labels. Then, the general idea of MLMix is to select the most appropriate hard label for mixed image. As illustrated in Fig. 2, MLMix involves three parameters, i.e., λ_x, λ_{y_i}, λ_{y_j}, and the augmented MLMix dataset is $\widetilde{\mathcal{D}} = (\widetilde{\mathcal{X}}, \widetilde{\mathcal{Y}}) = \{(\widetilde{x}_i, \widetilde{y}_i)\}_{i=1}^N$:

$$\begin{aligned} \widetilde{x}_i &= \lambda_x x_i + (1-\lambda_x)x_j, \\ \widetilde{y}_i &= \lambda_{y_i} y_i + \lambda_{y_j} y_j + (1-\lambda_{y_i}-\lambda_{y_j})(y_i \cup y_j), \end{aligned} \tag{2}$$

where λ_x is randomly drawn from the beta distribution, and λ_{y_i}, λ_{y_j} are weighting factors for selecting the correct hard label.

Given a mixed image with randomly sampled λ_x, interpolation policies of λ_{y_i} and λ_{y_j} are optimized using the meta-learning strategy (step 1–4 in Fig. 2), so as to obtain the compatible hard la-

bel for each mixed image and ensure the unbiased model optimization. The basic intuition behind this data-driven hard label derivation is that the optimal interpolation policies should target to minimize the segmentation loss on validation data (Zhang et al., 2020b) and promote the generalization ability of model. Specifically, in each training iteration, a meta net is copied from the original segmentation net, and a mini-batch of MLMix samples is forward passed through the meta net (step 1 in Fig. 2). Then, the parameters in the meta net are updated by moving the current $\mathbf{w}^t$ along the descent direction of segmentation loss as follows (step 2 in Fig. 2):

$$\hat{\mathbf{w}}^{t+1}(\lambda_{y_i}, \lambda_{y_j}) = \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \mathcal{L}(\widetilde{y}_i, f_{\mathbf{w}^t}(\widetilde{x}_i)), \tag{3}$$

where MLMix label $\widetilde{y}_i$ is a function of parameters λ_{y_i} and λ_{y_j} that are differentiable. $\mathcal{L}(\cdot)$ indicates the cross-entropy loss. Similar to well-known model-agnostic meta-learning (MAML) (Finn et al., 2017) with second-order back-propagation, we optimize the weighting factors λ_{y_i} and λ_{y_j} by minimizing the segmentation loss on validation data (i.e., original data without mixup) with the feedback from the updated parameters $\hat{\mathbf{w}}^{t+1}$ (step 3 and 4 in Fig. 2). In our implementation, we calculate the gradients of λ_{y_i}, λ_{y_j} and pick out the most compatible hard label by referring to the adjusted weighting factors $\widehat{\lambda_{y_i}}$, $\widehat{\lambda_{y_j}}$:

$$\begin{aligned}
\widehat{\lambda_{y_i}} &= \lambda_{y_i} - \eta \frac{\partial}{\partial \lambda_{y_i}} \mathbb{E}[\mathcal{L}(y_{val}, f_{\hat{\mathbf{w}}^{t+1}}(x_{val}))], \\
\widehat{\lambda_{y_j}} &= \lambda_{y_j} - \eta \frac{\partial}{\partial \lambda_{y_j}} \mathbb{E}[\mathcal{L}(y_{val}, f_{\hat{\mathbf{w}}^{t+1}}(x_{val}))],
\end{aligned} \tag{4}$$

where η is learning rate. (x_{val}, y_{val}) represents an image-label pair in validation set. Then, we select the maximum weighting factor (among $\widehat{\lambda_{y_i}}$, $\widehat{\lambda_{y_j}}$, $1-\widehat{\lambda_{y_i}}-\widehat{\lambda_{y_j}}$), and the corresponding hard label is regarded as the final MLMix label:

$$\widetilde{y}_i = \begin{cases} y_i, & \widehat{\lambda_{y_i}} \text{ is maximum;} \\ y_j, & \widehat{\lambda_{y_j}} \text{ is maximum;} \\ y_i \cup y_j, & \text{otherwise.} \end{cases} \tag{5}$$

Consequently, the optimal hard label is derived with respect to the validation data performance improvement. Different from the vanilla mixup obtaining soft label from the prior beta distribution, MLMix utilizes the second-order back-propagation to obtain the compatible hard label for each mixed image, thereby facilitating the generalization ability of segmentation net.

After the meta steps, each mixed image is equipped with an optimized hard label in an online fashion. Since the hard label learns from the meta-knowledge in the validation data, it can provide the consistent category information with clinical diagnosis, and the augmented data can be utilized for the unbiased training of segmentation net (step 5 in Fig. 2). The learning procedure of MLMix is summarized in Algorithm 1. As illustrated in Fig. 3 (d), our MLMix label provides precise category information for polyp segmentation.

3.2. Confidence-aware resampling (CAR) strategy

With the augmented and original datasets, the pixel-level cross-entropy loss and image-level Jaccard loss are widely utilized to optimize the segmentation models (Guo et al., 2021a). Specifically, the image-level loss between the predicted area and the ground truth polyp area is calculated by $\mathcal{L}_{img} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{img}^{i} = \frac{1}{N}\sum_{i=1}^{N}(1-\frac{p_i y_i}{p_i+y_i-p_i y_i})$, where $p_i = f_{\mathbf{w}}(x_i)$. (x_i, y_i) refers to ith polyp image and its label, and N is the total number of images. Define j as the position of pixel and M^c as the number of pixels in class c, then the pixel-level classification errors are calculated by $\mathcal{L}_{pix} = \frac{1}{N\times M}\sum_{c=1}^{C}\sum_{i=1}^{N}\sum_{j=1}^{M^c}\mathcal{L}_{pix}^{c,i,j} = -\frac{1}{N\times M}\sum_{c=0}^{C}\sum_{i=1}^{N}\sum_{j=1}^{M^c} y_{ij}^c \log p_{ij}^c$, where $M = M^0 + M^1$. $c=0$ indicates the normal class and $c=1$ represents the class of polyp. Then, the final objective function for segmentation can be represented as $\mathcal{L} = \mathcal{L}_{img} + \mathcal{L}_{pix}$. However, this objective function treats training images and pixels equally; thus, the learning procedure tends to become unstable with the variability of polyp images and the outliers in MLMix data. Thereby, the learning procedure gets stuck in the local optima. Moreover, the segmentation model may learn an insufficient representation in the minor class.

Algorithm 1 : Optimization.

Input: Training data $\mathcal{D} = (\mathcal{X}, \mathcal{Y}) = \{(x_i, y_i)\}_{i=1}^{N}$, validation data $\mathcal{D}_v = (\mathcal{X}_v, \mathcal{Y}_v) = \{(x_v, y_v)\}_{v=1}^{N_v}$
Parameters: Segmentation net ($\mathbf{w}$), λ_x, λ_{y_i}, λ_{y_j}

1: Initialize $\mathbf{w}$ with pre-trained model, randomly sample λ_x from beta distribution and initialize $\lambda_{y_i} = \lambda_{y_j} = \frac{1}{3}$
2: **for** t=1 to T **do**
3: Apply mixup on randomly selected pairs of images and obtain their corresponding labels with uniformly initialized $\lambda_{y_i}, \lambda_{y_j}$ via Eq. (2), which are denoted as $\widetilde{\mathcal{D}} = (\widetilde{\mathcal{X}}, \widetilde{\mathcal{Y}}) = \{(\widetilde{x}_i, \widetilde{y}_i)\}_{i=1}^{N}$.
4: Optimize the meta net ($\mathbf{w}^t$) on $\widetilde{\mathcal{D}}$ to $\hat{\mathbf{w}}^{t+1}$ via Eq. (3) (step 1, 2 in Figure 2)
5: Optimize weighting factors λ_{y_i} and λ_{y_j} on $\mathcal{D}_v$ via Eq. (4) (step 3, 4 in Figure 2)
6: Update MLMix label $\widetilde{y}_i$ via Eq. (5) to reconstruct $\widetilde{\mathcal{D}}$
7: Optimize the segmentation net ($\mathbf{w}^t$) on reconstructed $\{\mathcal{D}, \widetilde{\mathcal{D}}\}$ to $\mathbf{w}^{t+1}$ (step 5 in Figure 2)
8: **end for**
9: return $\mathbf{w} = \mathbf{w}^{T+1}$

To remedy these drawbacks, we propose CAR strategy that comprises an easy-to-hard gradual learning scheme and joints image- and pixel-level optimization for hierarchical learning and class rebalancing. For the image-level CAR, we follow (Li and Gong, 2017; Qin et al., 2020) and modify the objective function of $\mathcal{L}_{img}$ as

$$\min_{\mathbf{u},\mathbf{w}} \frac{1}{N}\left(\sum_{i=1}^{N} u_i \mathcal{L}_{img}^{i} - \lambda_{img}\sum_{i=1}^{N} u_i\right), \tag{6}$$

where $-\lambda_{img}\sum_{i=1}^{N} u_i$ is the self-paced regularizer. $\mathbf{u} = [u_1, u_2, \ldots, u_N] \in \mathbb{R}^N,\ s.t.\ u_i \in \{0,1\}$ are weights of training images, and λ_{img} is the age parameter that controls the number of selected images and regularizes the learning pace.

In the unselected complex images, there are pixels with high confidence scores, which also play an important role in the segmentation net optimization. Instead of merely selecting confident images in (Li and Gong, 2017; Qin et al., 2020), we additionally select pixels based on pixel-level loss in an easy-to-hard way to retrain the segmentation net, which is formulated as

$$\min_{\mathbf{v},\mathbf{w}} \frac{1}{N\times M}\left(\sum_{c=0}^{C}\sum_{i=1}^{N}\sum_{j=1}^{M^c} v_{ij}^c \mathcal{L}_{pix}^{c,i,j} - \sum_{c=0}^{C}\lambda_{pix}^c \sum_{i=1}^{N}\sum_{j=1}^{M^c} v_{ij}^c\right), \tag{7}$$

where $-\sum_{c=0}^{C}\lambda_{pix}^c\sum_{i=1}^{N}\sum_{j=1}^{M^c} v_{ij}^c$ is the self-paced regularizer. $\mathbf{v} = [v_{11}, v_{12}, \ldots, v_{NM-1}, v_{NM}] \in \mathbb{R}^{N\times M}, s.t.\ v_i \in \{0,1\}$ are weights of training pixels, and λ_{pix}^c is the age parameter that controls the amount of selected pixel in cth class. To rebalance the selected data distribution during the training process and alleviate the class imbalance problem, we incorporate the class prior knowledge by assigning different λ_{pix}^c for each class.

Since the loss derivation of the segmentation net often fluctuates at the early training stage, it is hard to manually set initial $\lambda_{img}, \lambda_{pix}$ as well as the corresponding increasing paces (Li and Gong, 2017). A feasible way is searching the solution path with respect to the number of images and pixels involved in training. Specifically, $\mathcal{L}_{img}$ and $\mathcal{L}_{pix}$ are firstly sorted across all images and pixels in ascending orders. Thus, we can select reliable images and pixels with high confidence according to the penalty proportion parameters r_{img}, r^0_{pix}, r^1_{pix}, which progressively vary from low to high. Further, we dynamically adjust the proportion parameters based on the segmentation performance, which is estimated from the overlap score Ac between predictions and labels in the training data. At the tth epoch, the proportion parameters are adjusted according to:

$$\begin{cases} r_{img} \mathrel{+}= \alpha, r^0_{pix} \mathrel{+}= \beta, r^1_{pix} \mathrel{+}= \gamma, & Ac > A_{t-1}; \\ r_{img}, r^0_{pix}, r^1_{pix}, & \text{otherwise}. \end{cases} \tag{8}$$

Only when the overlap score of the current epoch Ac surpasses the accumulated score A_{t-1} in the previous epochs, the proportion parameters will increase by predefined update steps (α, β, γ) to incorporate more credible images and pixels. Otherwise, the proportion parameters remain unchanged to prevent the segmentation net from crashing with overwhelming low-quality samples.

We alternatively optimize Eqs. (6) and (7) with respect to **u**, **v** and **w**. More specifically, the optimization strategy is summarized in Algorithm 2

Algorithm 2 : Optimization.

Input: Training data $\{\mathcal{D}, \widetilde{\mathcal{D}}\} = \{(x_i, y_i)\}_{i=1}^N$
Parameters: Segmentation net (**w**)

Initialize **w** and the accuracy of training data $A_0 = 0$
Initialize proportion parameters $r_{img} = 0.2$, $r^0_{pix} = r^1_{pix} = 0.1$
for t=1 to T **do**
  Implement MLMix to construct $\{\mathcal{D}, \widetilde{\mathcal{D}}\}$, which are utilized to train the model in current iteration.
  Update **u** and **v** via Eq. (9) and Eq. (10) to construct $\mathbb{S}_{img}$ and $\mathbb{S}_{pix}$.
  Update **w** in the segmentation model with resampled $\{\mathbb{S}_{img}, \mathbb{S}_{pix}\}$ via Eq. (11).
  Calculate Ac and the accumulated accuracy $A_t = 0.5 \times (Ac - A_{t-1}) + A_{t-1}$.
  Update $r_{img}, r^0_{pix}, r^1_{pix}$ according to Eq. (9).
end for
return $\mathbf{w} = \mathbf{w}^*$

3.2.1. Update **u** with **w** fixed

Using the alternating convex optimization strategy (Li and Gong, 2017), for fixed **w**, the closed-form solution of $\mathbf{u}^*$ can be easily calculated by

$$u^*_i = \begin{cases} 1, & \mathcal{L}^i_{img} \leqslant \lambda_{img}; \\ 0, & \text{otherwise}. \end{cases} \tag{9}$$

Intuitively, when updating the weight variable **u** with fixed **w**, an image is selected ($u^* = 1$) when its loss is smaller than λ_{img}.

3.2.2. Update **v** with **w** fixed

For fixed **w**, the closed-form solution v^* is

$$v^*_{ij} = \begin{cases} 1, & \mathcal{L}^{c,i,j}_{pix} \leqslant \lambda^c_{pix}; \\ 0, & \text{otherwise}. \end{cases} \tag{10}$$

Similarly, a pixel is selected ($v^* = 1$) when its loss is smaller than λ_{pix} of its corresponding class. However, if pixels from different classes are treated equally, model tends to be biased towards the background class and ignores some polyp regions. To tackle this issue, we propose to assign relatively larger λ^1_{pix} for polyp class and smaller λ^0_{pix} for background class; thus, the update step γ should be set to be larger than β. Therefore, CAR strategy can encode the class prior knowledge to balance data involved in the training procedure.

3.2.3. Update **w** with **u**, **v** fixed

After the eligible images and pixels are selected with the fixed **w**, the segmentation network is retrained using the selected set $\{\mathbb{S}_{img}, \mathbb{S}_{pix}\}$. Then, the objective function (the summation of Eqs. (6) and (7)) can be degenerated as the following,

$$\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} u^*_i \mathcal{L}^i_{img} + \frac{1}{N \times M} \sum_{c=0}^{C} \sum_{i=1}^{N} \sum_{j=1}^{M^c} v^{c*}_{ij} \mathcal{L}^{c,i,j}_{pix}. \tag{11}$$

It is obvious that Eq. (11) assigns different weights for the training images and pixels according to their confidences. Thus, we can resort to stochastic gradient descent (SGD) and gradient back-propagation methods to optimize parameters **w** in the segmentation net.

At the beginning of the learning stage, the model is unstable, and only confident images and pixels are involved to learn representative patterns of polyps. With the size of involved training samples increasing along with the learning procedure, the model is trained with more challenging patterns. This joint image- and pixel-level easy-to-hard gradual learning scheme leads to robust and discriminative feature extraction. Moreover, λ^c_{pix} for each class is assigned based on the class distribution prior knowledge to ensure the balance of the selected class distribution and alleviate the class imbalance problem. Under the alternating optimization of parameters (i.e., **u**, **v** **w**), the objective function in Eq. (11) can iteratively decrease to an optimal value. Thus, the segmentation net accelerates the learning convergence and becomes increasingly stable.

4. Experiments and results

4.1. Experimental setup

4.1.1. Datasets

Two polyp image datasets were utilized to verify the effectiveness of the proposed methods.

EndoScene dataset. This benchmark dataset (Vázquez et al., 2017) includes 912 colonoscopy images with corresponding pixel-wise annotations. We follow the standard setup in (Vázquez et al., 2017) with the constraint that images captured from one patient cannot be in different sets, and obtain 547 training, 183 validation, 182 test colonoscopy images.

WCE polyp dataset. Our private polyp dataset consists of 541 WCE images. They are collected through the Medtronics Pillcam WCE in the Prince of Wales Hospital. The ground truths of polyp areas were depicted by two professional experts. We randomly split this dataset for fourfold cross-validation and keep data preparation methods same across different experiments.

4.1.2. Implementation

The proposed method was implemented with the PyTorch library. DeepLabv3+ (Chen et al., 2018) was adopted as the backbone of segmentation net. ResNet-101 (He et al., 2016) was leveraged as the encoder and initialized with the pre-trained parameters obtained on ImageNet dataset. We utilized polynomial learning rate scheduling, where the initial learning rate is 0.001 and the maximum epoch number is 500. In each iteration, the batch size is empirically set as 8. The update steps of penalty proportion parameters α, β, and γ in CAR strategy were set to $\frac{2}{500}$, $\frac{1}{500}$, and

Table 1
Polyp segmentation results in comparison with state-of-the art methods.

Datasets	Methods	*Dice (%)*	*Jac (%)*	*Acc (%)*	*F2-score (%)*	*Sen (%)*	*Spe (%)*
EndoScene	Vázquez et al. (2017)	80.099	72.320	96.296	79.061	78.986	99.439
	Zhou et al. (2018)	79.842	71.756	95.885	80.252	82.492	98.622
	Fang et al. (2019)	81.987	73.654	96.438	81.859	82.630	99.306
	Qadir et al. (2019)	84.145	77.369	96.877	83.433	84.575	99.328
	Wickstrøm et al. (2020)	81.867	74.542	96.643	81.731	82.130	99.286
	Zhang et al. (2020a)	85.417	78.955	97.020	84.483	84.939	**99.507**
	Fan et al. (2020)	85.501	78.077	96.861	84.845	84.979	99.501
	Ours	**87.450**	**80.808**	**97.242**	**88.143**	**90.173**	98.904
WCE	Vázquez et al. (2017)	73.422±1.304	64.113±0.808	97.969±0.355	73.048±1.644	73.562±2.199	99.010±0.321
	Zhou et al. (2018)	80.811±1.791	72.326±1.426	98.265±0.337	80.808±1.850	81.273±1.916	99.165±0.170
	Fang et al. (2019)	75.106±0.662	65.130±0.881	97.118±0.574	74.556±0.626	74.960±0.877	99.306±0.132
	Qadir et al. (2019)	82.927±1.342	74.096±1.513	98.188±0.165	84.031±1.461	85.847±1.689	98.580±0.304
	Wickstrøm et al. (2020)	78.103±2.778	69.502±2.289	98.248±0.243	78.033±2.389	78.576±2.103	99.162±0.094
	Zhang et al. (2020a)	85.619±0.741	77.709±0.816	98.561±0.134	85.839±1.013	86.576±1.336	99.024±0.293
	Fan et al. (2020)	84.131±1.291	74.525±1.240	98.404±0.163	85.935±1.705	87.593±2.016	98.859±0.256
	Ours	86.453±1.070	78.007±1.188	98.562±0.118	89.222±1.090	91.941±0.975	98.619±0.185

$\frac{2}{500}$, respectively. To enrich the limited training samples, we employed online data augmentations, such as random rotation and crop. The augmented images were then resized to the resolution of 256×256 for training.

4.1.3. Evaluation metrics

The segmentation performance was evaluated by six commonly utilized metrics, i.e., *Dice*, Jaccard score (*Jac*), accuracy (*Acc*), *F2-score*, sensitivity (*Sen*) and specificity (*Spe*). For all the evaluation metrics, a higher score indicates a better segmentation performance. Best and second best results are **highlighted** and underlined.

4.2. Results on EndoScene dataset

We first assessed the performance of the proposed approach (*row* 7) and compared it with state-of-the-art polyp segmentation methods (Vázquez et al., 2017; Zhou et al., 2018; Wickstrøm et al., 2020; Fang et al., 2019; Qadir et al., 2019; Zhang et al., 2020a; Fan et al., 2020). The comparison results on the EndoScene dataset are listed in Table 1. For a fair comparison, we implemented their network architectures and utilized the same online data preparation methods. It is observed that the proposed method shows superior performance with increments of 7.351%, 7.608%, 5.463%, 3.305%, 5.583%, 2.033%, 1.949% in *Dice*, 11.187%, 7.681%, 7.543%, 5.598%, 8.043%, 5.234%, 5.194% in *Sen* compared with methods (Vázquez et al., 2017; Zhou et al., 2018; Fang et al., 2019; Qadir et al., 2019; Wickstrøm et al., 2020; Zhang et al., 2020a; Fan et al., 2020), respectively. Among the evaluation metrics, *Sen* score indicates the proportion of polyp pixels that are accurately identified, which is critical in the clinical practice. Since the area of polyp region is smaller than that of normal one, previous methods (Vázquez et al., 2017; Zhou et al., 2018; Wickstrøm et al., 2020; Fang et al., 2019; Qadir et al., 2019; Zhang et al., 2020a; Fan et al., 2020) are prone to categorize the polyp pixel as normal one, which usually leads to a poor sensitivity and a relatively high specificity score. On the contrary, the significant improvement in *Sen* score reveals that the proposed method has an inherent ability of tackling the class imbalance problem.

We visualized four typical polyp images and compared the corresponding segmentation predictions of methods (Vázquez et al., 2017; Zhou et al., 2018; Wickstrøm et al., 2020; Fang et al., 2019; Qadir et al., 2019; Zhang et al., 2020a; Fan et al., 2020) and our method in Fig. 4. It is obvious that existing state-of-the-art methods under-segment regions with low contrast characteristics (*rows* $1-2$ in Fig. 4) and different illumination conditions (*row* 3 in Fig. 4). In contrast, the proposed method can make accurate predictions in those error-prone regions. Additionally, our model is more resistant against specular reflection and outperforms other methods in manifested reflection regions (*row* 4 in Fig. 4).

4.3. Results on WCE polyp dataset

We then performed the proposed MLMix data augmentation method and CAR strategy on the WCE polyp dataset, and exhibited the corresponding segmentation results with mean and standard deviation of evaluation metrics, as illustrated in Table 1. It is obvious that the proposed method shows superior segmentation performance compared with other polyp segmentation methods (Vázquez et al., 2017; Zhou et al., 2018; Fang et al., 2019; Qadir et al., 2019; Wickstrøm et al., 2020; Zhang et al., 2020a; Fan et al., 2020) with increments of 13.031%, 5.642%, 11.347%, 3.526%, 8.350%, 0.834%, 2.322% in *Dice* score. Fig. 4 (*rows* $5-8$) visualizes segmentation predictions of WCE images for the purpose of quantitative comparison. Due to the relatively low resolution of WCE images, boundaries of polyp regions are usually blurred (*rows* $4-5$), and a high degree of apparent similarity share between polyp and normal tissues (*rows* $6-7$). In this scenario, other methods are incapable of identifying polyp regions accurately and result in missing detections, while our approach still performs well. Both quantitative and qualitative evaluations on two polyp datasets demonstrate the superiority of the proposed method.

4.4. Ablation study of MLMix

To analyze the effectiveness of the proposed MLMix data augmentation method, we conducted ablation experiments to compare it with other mixup related data augmentation methods, including mixup (Zhang et al., 2018), asym. mixup (asymmetric mixup) (Li et al., 2019), CGMMix (Guo et al., 2021a). Mixup conducts data augmentation through convex combination on images and derives soft labels. Asym. mixup further applies threshold to obtain hard label for medical image segmentation. Our previous CGMMix considers the varying degrees of CRC and incorpotates a confidence-guided manifold mixup in both image and feature levels. The comparison results recorded in *rows* $3-6$ of Table 2 demonstrate that the proposed MLMix performs favorably against other data augmentation methods (Zhang et al., 2018; Li et al., 2019; Guo et al., 2021a). In particular, MLMix achieves a prominent *Jac* of 78.88%, which shows increments of 2.31%, 1.44%, 0.47% in comparison to (Zhang et al., 2018; Li et al., 2019; Guo et al., 2021a), respectively. This result reveals the good capability of MLMix to enrich the limited training dataset for polyp segmentation. Moreover, the im-

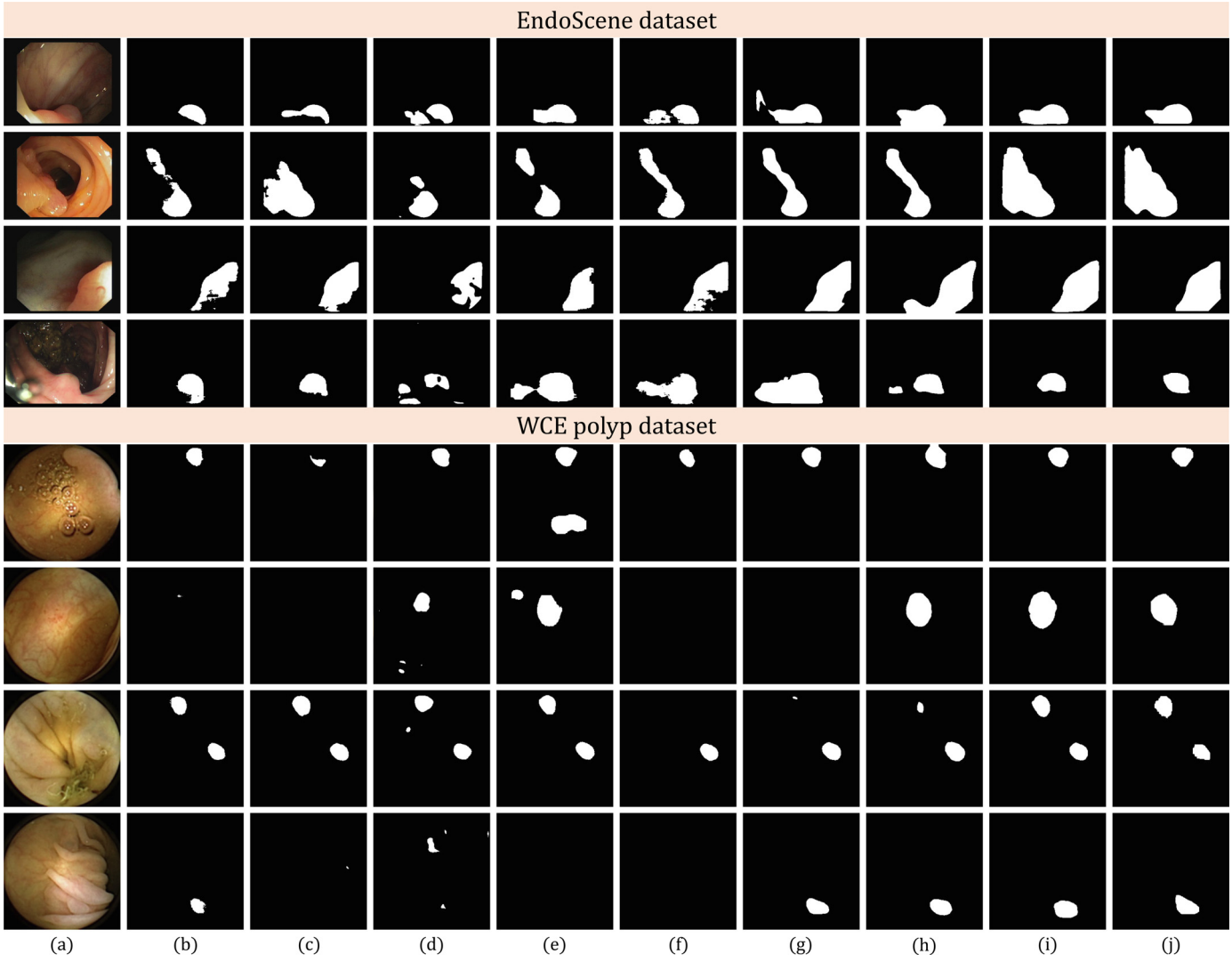


Fig. 4. Typical examples of segmentation results on the EndoScene dataset (*rows* 1 – 4) and the WCE polyp dataset (*rows* 5 – 8). Each row presents (a) input image, segmentation predictions of models (b) FCN (Vázquez et al., 2017), (c) UNet++ (Zhou et al., 2018), (d) SFA (Fang et al., 2019), (e) Mask R-CNN (Qadir et al., 2019), (f) SegNet (Wickstrøm et al., 2020), (g) ACSNet (Zhang et al., 2020a), (h) PraNet (Fan et al., 2020), (i) ours and (j) ground truth.

Table 2
Ablation studies for MLMix on EndoScene.

Methods	*Dice*	*Jac*	*Acc*	*F2*	*Sen*	*Spe*
Baseline	82.52	74.93	96.44	82.02	82.30	99.31
w/ Overlap mixup	83.07	75.14	96.52	83.00	83.71	99.16
w/ Mixup (Zhang et al., 2018)	83.55	76.51	96.64	83.08	83.60	99.39
w/ Asym. mixup (Li et al., 2019)	84.69	77.44	96.73	84.63	85.42	99.33
w/ CGMMix (Guo et al., 2021a)	85.43	78.41	96.89	85.24	85.79	**99.45**
w/ MLMix (Ours)	85.92	78.88	96.94	85.70	86.33	99.35
Ours	**87.45**	**80.81**	**97.24**	**88.14**	**90.17**	98.90

provements of MLMix in comparison to baseline (*row* 2) and overlap mixup (*row* 3, mixed hard label is $y_i \cup y_j$) demonstrate that incorporating meta-learning strategy to learn the data-driven interpolation policy enables the segmentation model to generate compatible hard label for mixed images. It is worth noting that the proposed MLMix with CAR strategy (*row* 7) significantly boosts the segmentation performance of the baseline model, DeepLabv3+ (Chen et al., 2018), and achieves a *Dice* of 87.45% and a *Sen* of 90.17%. The remarkable improvements of 4.93% in *Dice* and 7.87% in *Sen* compared with the baseline model validate the proposed MLMix and CAR strategy contribute to the performance gains.

4.5. Ablation study of CAR strategy

To deal with the variability of polyp images and class imbalance problems, two typical strategies, HLS (hierarchical learning strategy) (Qin et al., 2020) and asym. focal loss (asymmetric focal loss) (Li et al., 2019), are commonly utilized, as shown in Table 3. HLS emphasizes confident images to tackle the issue of variable training instances and prevent the negative effect of outliers, while asym. focal loss penalizes uncertain instances to alleviate the class imbalance problem. Our CAR strategy integrates the confidence inference and leverages class prior knowledge to simultaneously deal with the aforementioned two data biased problems in polyp seg-

Table 3
Ablation studies for CAR strategy on EndoScene.

Methods	*Dice*	*Jac*	*Acc*	*F2*	*Sen*	*Spe*
Baseline	82.52	74.93	96.44	82.02	82.30	99.31
w/ HLS (Qin et al., 2020)	84.72	77.65	96.81	84.41	85.07	**99.39**
w/ Asym. focal loss (Li et al., 2019)	84.23	77.04	96.76	84.52	85.33	99.33
w/ CAR (Ours)	86.09	79.26	97.04	86.20	87.49	99.35
Ours	**87.45**	**80.81**	**97.24**	**88.14**	**90.17**	98.90

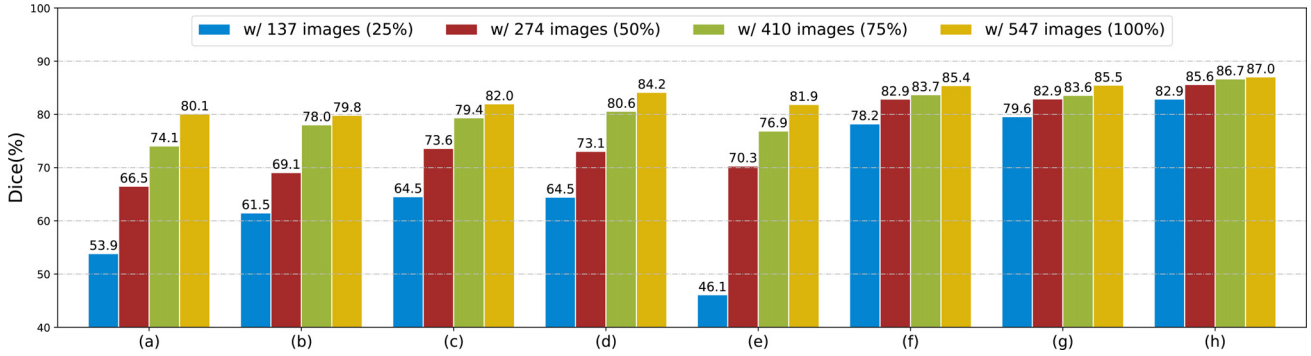


Fig. 5. Comparison results of segmentation models trained with different numbers of training images. From left to right, segmentation results are obtained from models (a) FCN (Vázquez et al., 2017), (b) UNet++ (Zhou et al., 2018), (c) SFA (Fang et al., 2019), (d) Mask R-CNN (Qadir et al., 2019), (e) SegNet (Wickstrøm et al., 2020), (f) ACSNet (Zhang et al., 2020a), (g) PraNet (Fan et al., 2020), (h) ours. Note that "274 (50%)" indicates 274 images (50 percent of EndoScene training set) are involved for optimization.

mentation and shows prominent increments of 1.61%, 2.22% in *Jac* compared with other reweighting methods, HLS and asym. focal loss methods.

Fig. 6 plots the *Dice* and *Jac* curves on the test data, under different methods (Baseline, HLS Qin et al., 2020, asym. focal loss Li et al., 2019 and the proposed CAR strategy), where the x-axis denotes the training epoch. The figure shows two insights. First, *Dice* and *Jac* curves are stable at the later stages of training procedure. This empirically verifies the convergence of the segmentation models with different strategies. Second, the easy-to-hard learning scheme is demonstrated to have capability of accelerating the convergence since both HSL and CAR strategies converge to a steady state slightly faster than other methods. With the incorporation of pixel-level easy-to-hard gradual learning, our CAR strategy is superior than HSL that only considers image-level hierarchical learning strategy.

We then illustrate the selected images and pixels at different training stages to visualize the learning schemes of CAR strategy, as shown in Fig. 7. It is observed that at the early learning stage, only a small set of confident images and pixels are involved to learn representative patterns of polyps. Notably, the involved pixels of foreground and background classes are relatively balanced in comparison to class distribution of original data. Along with the learning procedure, more challenging images and pixels are introduced to train the segmentation. Moreover, since there exist pixels with high confidence scores in those unselected complex images, the pixel-level easy-to-hard learning scheme is demonstrated to be complementary to the image-level one within the proposed CAR strategy.

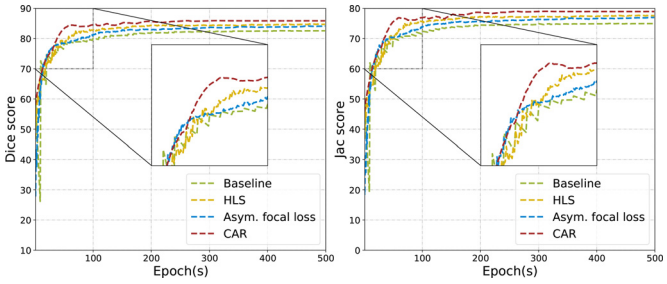


Fig. 6. *Dice* and *Jac* curves on EndoScene test data w.r.t. training process.

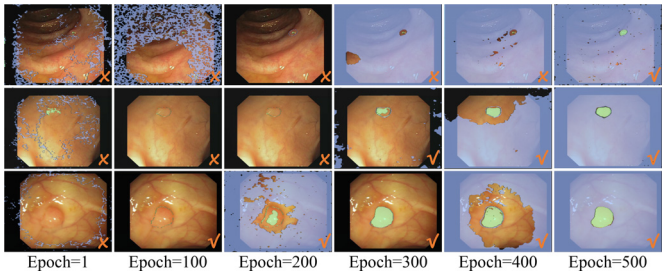


Fig. 7. Illustration of the selected images (orange check marks), foreground pixels (green regions), and background ones (blue regions) w.r.t. training process (i.e., at epoch 1, 100, 200, 300, 400, 500) with the proposed CAR strategy. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.6. Different numbers of training images

Sufficient annotated data is crucial for the optimization of deep CNNs, and can promote the generalization capability of the optimized segmentation model. To further verify the effectiveness of the proposed method, we compared it with state-of-the-art polyp segmentation models (Vázquez et al., 2017; Zhou et al., 2018; Wickstrøm et al., 2020; Fang et al., 2019; Qadir et al., 2019; Zhang et al., 2020a; Fan et al., 2020) optimized with different numbers of training images. The numbers of training images ranged from 25% to 100% of the total training set (547 images in EndoScene dataset) in increments of 25% proportion, and Fig. 5 shows the segmentation performance in terms of the *Dice* score. In general, it is clear that our method is more stable and consistently performs superior than other methods with different training images, validating the robustness of the proposed method. When the training data is scarce (137 images), other segmentation models exhibit a sharp decline in *Dice* score with unsatisfactory generalization performance. On the contrary, the proposed method shows a favorable performance and achieves significant promotions in comparison to

Table 4
Performance of our method on EndoScene dataset with different segmentation baseline networks. Best results are **highlighted**.

Methods	*Dice*	*Jac*	*Acc*	*F2*	*Sen*	*Spe*
UNet (15′) (Ronneberger et al., 2015)	76.07	67.12	95.63	75.21	75.76	99.05
Ours (UNet)	**83.20**	**74.77**	**96.64**	**83.96**	**85.25**	**99.09**
SegNet (17′) (Badrinarayanan et al., 2017)	81.87	74.54	96.64	81.73	82.13	99.29
Ours (SegNet)	**84.10**	**76.76**	**96.82**	**84.81**	**85.96**	**99.30**
MultiResUNet (20′) (Ibtehaz and Rahman, 2020)	81.31	73.94	96.62	80.24	80.14	**99.46**
Ours (MultiResUNet)	**83.36**	**76.26**	**96.66**	**83.65**	**84.73**	99.01
PraNet (20′) (Fan et al., 2020)	85.50	78.08	96.86	84.85	84.98	**99.50**
Ours (PraNet)	**86.77**	**80.08**	**97.32**	**87.91**	**90.74**	98.86
CS2-Net (20′) (Mou et al., 2021)	80.06	71.85	96.20	79.55	79.94	99.31
Ours (CS2-Net)	**82.89**	**75.89**	**96.70**	**82.75**	**83.41**	**99.35**
DeepLabv3+ (18′) (Chen et al., 2018)	82.52	74.93	96.44	82.02	82.30	**99.31**
Ours (DeepLabv3+)	**87.45**	**80.81**	**97.24**	**88.14**	**90.17**	98.90

methods (Vázquez et al., 2017; Zhou et al., 2018; Fang et al., 2019; Qadir et al., 2019; Wickstrøm et al., 2020; Zhang et al., 2020a; Fan et al., 2020) with 29.0%, 21.4%, 18.4%, 18.4%, 36.8%, 4.7%, 3.3% increments in *Dice* score. This observation demonstrates the generalization capacity of our approach, and it is highly promising in real clinical practice where annotated datasets are scarce.

4.7. Comparisons with different backbones

The proposed MLMix and CAR strategy methods were performed with the backbone of DeepLabv3+ (Chen et al., 2018) in previous experiments. To explore the generalization capacity of the proposed methods, we then integrated MLMix and CAR strategy to other segmentation backbones, including UNet (Ronneberger et al., 2015), SegNet (Badrinarayanan et al., 2017), MultiResUNet (Ibtehaz and Rahman, 2020), PraNet (Fan et al., 2020) and CS2-Net (Mou et al., 2021). Table 4 summarizes the comparison results with different backbone networks. It is observed that our method can greatly facilitate the performance over different baseline networks, UNet (Ronneberger et al., 2015), SegNet (Badrinarayanan et al., 2017), MultiResUNet (Ibtehaz and Rahman, 2020), PraNet (Fan et al., 2020), CS2-Net (Mou et al., 2021), DeepLabv3+ (Chen et al., 2018), with increments of 7.65%, 2.22%, 2.32%, 2.00%, 4.04%, 5.88% in *Jac* and 9.49%, 3.83%, 4.59%, 5.76%, 3.20%, 7.87% in *Sen*, respectively. The promising promotions reveal that the proposed method is general and could be integrated to existing segmentation models for polyp segmentation.

5. Conclusion

Automatic polyp segmentation is challenging due to the lack of large annotated datasets, the variability of polyps, and the class imbalanced data distribution. In this paper, we propose an MLMix data augmentation method and a CAR strategy to tackle the aforementioned issues. MLMix utilizes the meta-learning strategy to augment the limited training data and yield compatible image-label pairs in a data-driven manner. Further, the proposed CAR strategy adopts an easy-to-hard gradual learning scheme at both image and pixel levels, and leverages the class prior knowledge to balance the selected class distribution. The comprehensive experiments demonstrate the superiority of the proposed methods, which inherently can be transferred to extensive medical image segmentation baselines for data augmentation purpose and facilitating the robust feature extraction.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Xiaoqing Guo: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – review & editing. **Zhen Chen:** Conceptualization, Methodology, Writing – review & editing. **Jun Liu:** Conceptualization, Methodology, Investigation. **Yixuan Yuan:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62001410), and Hong Kong RGC Collaborative Research Fund grant C4063-18G (CityU 8739029).

References

Akbari, M., Mohrekesh, M., Nasr-Esfahani, E., Soroushmehr, S.R., Karimi, N., Samavi, S., Najarian, K., 2018. Polyp segmentation in colonoscopy images using fully convolutional network. In: International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 69–72.
Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.
Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A., 2019. MixMatch: a holistic approach to semi-supervised learning. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 5050–5060.
Cai, Q., Pan, Y., Wang, Y., Liu, J., Yao, T., Mei, T., 2020. Learning a unified sample weighting network for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14173–14182.
Chaitanya, K., Karani, N., Baumgartner, C.F., Becker, A., Donati, O., Konukoglu, E., 2019. Semi-supervised and task-driven data augmentation. In: International Conference on Information Processing in Medical Imaging (IPMI), pp. 29–41.
Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., Zhou, Y., 2021. TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.
Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European Conference on Computer Vision (ECCV), pp. 801–818.
Fan, D.-P., Ji, G.-P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L., 2020. PraNet: parallel reverse attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 263–273.
Fang, Y., Chen, C., Yuan, Y., Tong, K.-y., 2019. Selective feature aggregation network with area-boundary constraints for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 302–310.
Finn, C., Abbeel, P., Levine, S., 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning (ICML), pp. 1126–1135.
Guo, X., Yang, C., Liu, Y., Yuan, Y., 2021. Learn to threshold: thresholdnet with confidence-guided manifold mixup for polyp segmentation. IEEE Trans. Med. Imaging 40 (4), 1134–1146.
Guo, X., Yang, C., Yuan, Y., 2021. Dynamic-weighting hierarchical segmentation network for medical images. Med. Image Anal. 102196.
Guo, X., Yuan, Y., 2019. Triple ANet: Adaptive abnormal-aware attention network for WCE image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 293–301.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.

Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B., 2020. AugMix: a simple method to improve robustness and uncertainty under data shift. In: International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=S1gmrxHFvB

Ibtehaz, N., Rahman, M.S., 2020. MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Netw. 121, 74–87.

Jha, D., Smedsrud, P.H., Johansen, D., de Lange, T., Johansen, H., Halvorsen, P., Riegler, M., 2021. A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation. IEEE J. Biomed. Health Inform..

Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D., 2019. ResUNet++: an advanced architecture for medical image segmentation. In: IEEE International Symposium on Multimedia (ISM), pp. 225–2255.

Jia, X., Mai, X., Cui, Y., Yuan, Y., Xing, X., Seo, H., Xing, L., Meng, M.Q.-H., 2020. Automatic polyp recognition in colonoscopy images using deep learning and two-stage pyramidal feature prediction. IEEE Trans. Autom. Sci. Eng. 17 (3), 1570–1584.

Jia, X., Xing, X., Yuan, Y., Xing, L., Meng, M.Q.-H., 2019. Wireless capsule endoscopy: a new tool for cancer screening in the colon with deep-learning-based polyp recognition. Proc. IEEE 108 (1), 178–197.

Jiang, L., Zhou, Z., Leung, T., Li, L.-J., Fei-Fei, L., 2018. MentorNet: learning data-driven curriculum for very deep neural networks on corrupted labels. In: International Conference on Machine Learning (ICML), pp. 2304–2313.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 1097–1105.

Li, H., Gong, M., 2017. Self-paced convolutional neural networks. In: International Joint Conference on Artificial Intelligence (IJCAI), pp. 2110–2116.

Li, Y., Vasconcelos, N., 2020. Background data resampling for outlier-aware classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13218–13227.

Li, Z., Kamnitsas, K., Glocker, B., 2019. Overfitting of neural nets under class imbalance: analysis and improvements for segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 402–410.

Lin, D., Li, Y., Nwe, T.L., Dong, S., Oo, Z.M., 2020. RefineU-Net: improved U-Net with progressive global feedbacks and residual attention guided local refinement for medical image segmentation. Pattern Recognit. Lett. 138, 267–275.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988.

Liu, X., Guo, X., Liu, Y., Yuan, Y., 2021. Consolidated domain adaptive detection and localization framework for cross-device colonoscopic images. Med. Image Anal. 102052.

Mou, L., Zhao, Y., Fu, H., Liu, Y., Cheng, J., Zheng, Y., Su, P., Yang, J., Chen, L., Frangi, A.F., et al., 2021. CS2-Net: deep learning segmentation of curvilinear structures in medical imaging. Med. Image Anal. 67, 101874.

Nguyen, N.-Q., Vo, D.M., Lee, S.-W., 2020. Contour-aware polyp segmentation in colonoscopy images using detailed upsamling encoder-decoder networks. IEEE Access 8, 99495–99508.

Qadir, H.A., Shin, Y., Solhusvik, J., Bergsland, J., Aabakken, L., Balasingham, I., 2019. Polyp detection and segmentation using mask R-CNN: does a deeper feature extractor CNN always perform better? In: International Symposium on Medical Information and Communication Technology (ISMICT), pp. 1–6.

Qin, W., Xiao, Z., Xie, Y., Yuan, Y., 2020. Self-paced learning for automatic prostate segmentation on mr images with hierarchical boundary sensitive network. In: IEEE International Conference on Real-time Computing and Robotics (RCAR), pp. 321–326.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241.

Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A., 2021. Cancer statistics, 2021. CA Cancer J. Clin. 71 (1), 7–33.

Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A., 2017. A benchmark for endoluminal scene segmentation of colonoscopy images. J. Healthc. Eng. 2017, 1–9.

Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., Bengio, Y., 2019. Manifold mixup: better representations by interpolating hidden states. In: International Conference on Machine Learning (ICML), pp. 6438–6447.

Wang, Q., Li, W., Gool, L.V., 2019. Semi-supervised learning by augmented distribution alignment. In: IEEE International Conference on Computer Vision (ICCV), pp. 1466–1475.

Wang, Z., Hu, G., Hu, Q., 2020. Training noise-robust deep neural networks via meta-learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4524–4533.

Wickstrøm, K., Kampffmeyer, M., Jenssen, R., 2020. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. Med. Image Anal. 60, 101619.

Wu, H., Zhong, J., Wang, W., Wen, Z., Qin, J., 2021. Precise yet efficient semantic calibration and refinement in convnets for real-time polyp segmentation from colonoscopy videos. In: The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI).

Yang, C., Guo, X., Zhu, M., Ibragimov, B., Yuan, Y., 2021. Mutual-prototype adaptation for cross-domain polyp segmentation. IEEE J. Biomed. Health Inform..

Yang, X., Wei, Q., Zhang, C., Zhou, K., Kong, L., Jiang, W., 2020. Colon polyp detection and segmentation based on improved MRCNN. IEEE Trans. Instrum. Meas. 70, 1–10.

Yuan, Y., Qin, W., Ibragimov, B., Han, B., Xing, L., 2018. RIIS-DenseNet: rotation-invariant and image similarity constrained densely connected convolutional network for polyp detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 620–628.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2018. mixup: Beyond empirical risk minimization. In: International Conference on Machine Learning (ICML). https://openreview.net/forum?id=r1Ddp1-Rb

Zhang, R., Li, G., Li, Z., Cui, S., Qian, D., Yu, Y., 2020. Adaptive context selection for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 253–262.

Zhang, Z., Zhang, H., Arik, S.O., Lee, H., Pfister, T., 2020. Distilling effective supervision from severe label noise. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9294–9303.

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2018. Unet++: a nested U-Net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA), pp. 3–11.